Clusters and Copies

An Analysis of Cryptocurrency Investment Scam Websites

Irene Klom

Delft University of Technology

TU Delft

Clusters and Copies

An Analysis of Cryptocurrency Investment Scam Websites

by

Irene Klom

to obtain the degree of Master of Science
Software Technology Track
with a 4TU specialization in Cyber Security
at the Delft University of Technology,
to be defended publicly on Wednesday, December 11, 2024 at 13.00.

Student number:	4663756	
Thesis committee:	Dr. R.S. (Rolf) van Wegberg MSc	Daily supervisor
	Dr. H.J. (Harm) Griffioen MSc	CS supervisor
	Daniël	FIOD supervisor
Project Duration:	March 2024 - December 2024	
Defense date:	December 11, 2024	
Faculty:	Faculty of Electrical Engineering, Mathematics and Computer Science, TU Delft	

Style:	TU Delft Report Style, with modifications by Daan Zwaneveld
Cover:	AI generated with DALL·E 3

TU Delft

Preface

This thesis is the final requirement to obtain my master's degree of Computer Science - Cyber Security. After a bachelor's and master's at the TU Delft, this university and city genuinely feel like home. I have learned so much, not the least about myself, and I have made friends for life. I feel privileged that I have been able to devote my full attention to my education for so long and I am proud to present this final document.

This master's thesis has taken me on an adventure filled with ups, downs, turns, and twists, but it has come to an end. Now that I reach the finish line, I find myself reflecting on the last nine months and the rewarding journey it has been. The time has flown by, and while the path wasn't always smooth, I can genuinely say this experience is one I will cherish for a lifetime. I feel ready to say goodbye for now to the academic environment, equipped with new knowledge, skills and experience. I cannot wait to find out what my next chapters in cyber security have in store for me.

I want to thank my daily supervisors, Rolf and Daniël, for providing me with the best guidance I could have wished for. To Rolf, I am happy to have experienced your enthousiasm for the field of cyber crime and I am grateful for your optimistic mentoring style. Daniël, I want to thank you for making me feel so very welcome at the FIOD, for sharing your experience and for all our weekly talks — always pleasant, never short. Many thanks Harm, for keeping an eye on the entire process and for providing me with valuable feedback. I would like to express my gratitude to my internship colleagues at the Financial Advanced Cyber Team of the FIOD for sharing all their experience in cyber investigations with me.
I won't forget all your fun stories and lunch conversations.

The final thanks is to my parents, my sister, my boyfriend, and to all of my friends and family.
You were always there supporting me and I know you always will be.

Irene Klom
Delft, December 2024

Summary

Cryptocurrency investment scams have become an increasingly prevalent threat, leveraging sophisticated methods to deceive and exploit victims. The phenomenon of *pig butchering* has gained prominence, but victims rarely see the perpetrators getting justice. There is a need for more decisive action of law enforcement, but several factors limit the feasibility of prosecution. In order to strengthen the knowledge position of law enforcement, this research aims to capture the Tactics, Techniques, and Procedures (TTP) of the owners and creators of cryptocurrency investment scam websites. The feasibility of conducting a criminal investigation into a scam operation is assessed. This feasibility is compared to that of a Notice and Takedown procedure. This is done by focusing on clusters of scam websites with very similar content. We created a dataset consisting of 436 websites. Each website belongs to one of four identified scam website templates. By monitoring the websites in the dataset over a period of 40 days, it was discovered that the majority of the websites goes offline after exactly one year. The contents of the webpages in the dataset are scraped and the similarities between the websites are measured. It was found that instances of the same template are similar enough, allowing us to use known scams to automatically detect new scam websites. Until generative AI becomes more widely used by scam website creators to fill their templates, fingerprints of known templates can be used to discover new templates. Clustering the websites based on their contents can identify which websites are likely to follow identical scamming procedures and which websites deviate from this. This can inform law enforcement on how to funnel investigative resources. The study was not able to conclude whether a template has one owner with multiple scammers hosting their own acquired instance(s) or that a template owner themself hosts multiple instances – with the existence of some copycats. There was evidence for both scenarios. The process of gathering metadata and content data revealed that free Cloudflare services provide scammers with additional security. Firstly, the Cloudflare Turnstile page prevents the website from being archived, which causes a lack of evidence about the website's content at the time of a scam. Secondly, Cloudflare's function as proxy hides the website's IP address, which creates an accountability void due to the absence of a geolocation. This complicates the process of starting a criminal investigation. In order for law enforcement to make a difference, it is advised to invest effort and money into creating an automated Notice and Takedown pipeline. This will reduce the effort needed to take down a scam and will simultaneously increase the effort and financial means needed by scam operations.

Contents

1

Introduction

The first cryptocurrency Bitcoin launched in 2009 [1]. Since then, its popularity and monetary value have increased significantly. Currently, more than 1,1 billion Bitcoin transactions have been completed and Bitcoin reached a record high value of over $99.000 per Bitcoin in November 2024 [2]. It has a market capitalization of over a trillion dollars [3]. Numerous other cryptocurrencies followed. The high popularity caused scammers to take advantage of the growing crypto market [4], [5].

According to Chainalysis [6], cryptocurrency scams had a revenue of at least $4.6 billion in 2023. This is the lowest reported number in the last 5 years, yet it is still “one of the biggest drivers of cryptocurrency-based crime” [6, p.104].

The average loss of a romance scam, a scam in which the scammer poses as a romantic interest and persuades the victim into investing in a scam, is $4.593, which is a tenfold of the average investment scam loss [6]. Although investment scams only make up a small portion of all cryptocurrency scams, victims categorized as victims of a romance scam, also known as pig butchering, will often lose their money to similar investment websites. Thus, the total losses to fake cryptocurrency investment websites and fake cryptocurrency exchange websites is more than just that of cases reported as investment scams. Developing new insights into investment scam websites can help to prevent people from falling victim to scammers.

One measure to combat scam websites is the Notice and Takedown (NTD) procedure. This Notice and Takedown procedure is the result of industry agreements and the Digital Services Act of the European Union [7]. The goal of the procedure is to provide a framework for taking illegal content offline and to provide more certainty about the legal liabilities of internet intermediaries, such as hosting providers. In the current situation, anyone can send a Notice and Takedown request to inform the hosting provider of a fraudulent website and to request the takedown of the website. The hosting providers must then review the takedown request and determine whether there is sufficient cause to take specific content offline or even to completely discontinue the hosting. As long as the hosting provider cooperates in the takedown, they will not be prosecuted for any illegal activities on websites that are hosted by their customers. Therefore, hosters have a strong incentive to comply with any NTD, especially those sent by law enforcement. It is thus important that law enforcement only sends NTD requests when there is sufficient evidence of wrongdoing [8].

Alternatively, if there is sufficient cause, law enforcement can decide to prosecute the owner of the website. However, this does not happen often, as there are several challenges in getting to the prosecution phase.

Current Dutch penal codes have a high threshold for proving a scam. If a vendor or service provider does not fulfill their promise, then this only qualifies as a scam if they never had the intention of fulfilling it in the first place. Therefore, law enforcement has to prove the malicious intent, which is a difficult task [9].

Victims of scams can sometimes not accurately recognize what happened exactly, which is another obstacle for law enforcement. As Cross describes: “Investment fraud victims initially struggle to identify that they have been deceived, as they instead perceive the loss to be a bad investment” [10, p. 7]. This can cause some victims to never realize that the situation is worth reporting. If victims have been scammed over a longer period of time and realize this only later on, then they might not recall the details of the scam at the moment of reporting.

Additionally, the difficulty of recognizing the difference between a scam and a legitimate matter, which has caused the person to fall victim to the scam, negatively influences the reliability of the witness statement in the police report. An example of this is when a victim claims that a famous person has promised them a high return on their investment, instead of claiming that someone posing as a famous person promised this. Because of this, the technical evidence of how a scam has occurred is valuable information in a fraud investigation.

Gathering evidence is a difficult task in a rapidly changing online environment. It can be even more difficult if the scam website has already gone offline at the moment of investigation. However, even if the website is online, then contents of the website might have changed over time. Internet archives can be leveraged to collect the historical data of a website.Not every website is archived equally often, since it mostly relies on snapshot requests made by users of the archive.

Unfortunately, not every crime can get investigated to the same extend. Due to the limited investigation capacity of law enforcement, solving larger crimes is prioritized over solving smaller ones. This means that the investigation of a scam website only becomes interesting when the total monetary value of the lost capital is significant enough. However, even when this is the case and evidence of this happening is secured, conviction of the perpetrator can still be challenging. It is hard to attribute the malicious website to an individual and it might be that the geolocation of the scammer(s) makes it impossible to move forward with the case.

The high threshold of proving malicious intent, the difficulties concerning reliable police reports, the challenges in securing sufficient technical evidence and the problems surrounding the attribution of the scam to individuals within a specific jurisdiction all cause the likelihood of eventually convicting the owners of scam websites to be very low.

This study aims to create a deeper insight into the phenomenon of cryptocurrency investment scam websites. It aims to combine the knowledge from previous studies with newly acquired knowledge from observing, measuring and analyzing the phenomenon of cryptocurrency scam websites in a quantitative manner. This should create novel insights into the tactics, techniques, and procedures (TTP) of cryptocurrency scammers. According to Bianco’s pyramid of pain, discovering TTP will have the most potential on disrupting the adversary’s business model [11].

Furthermore, this study can be used by law enforcement to inform decisions surrounding the feasibility of prosecution and to create more efficient tooling to discover scam websites, to gather technical evidence, and to take a website offline. Collectively, this study should strengthen the information position of law enforcement with regard to cryptocurrency investment scam websites.

To facilitate reproducibility and further research, the code and datasets used in this study are publicly accessible on GitHub[1].

[1] https://github.com/IreneKl/clusters_and_copies

2

Background

This chapter will introduce the concept of cryptocurrencies and cryptocurrency exchanges and will provide an insight into the most common scams. It will highlight the multiple entities that have a role in the prevention of online scams.

2.1. Cryptocurrencies

Cryptocurrencies are virtual currencies that rely on blockchain technology. The idea behind cryptocurrencies is that they use decentralized networks for transactions, which means that there is no central authority, such as a bank, controlling the transactions. All transactions are verified by a network and the distributed ledger is transparent and verifiable to everyone. This system works as long as more than half of the network is honest and does not actively undermine the system [1].

Even though all transactions are public, cryptocurrency wallets can be pseudonymously owned. This pseudonymity and the lack of a central authority make using cryptocurrencies attractive to criminals. Transactions have no upper limit and can happen in an instant. A reason not to use cryptocurrencies as a criminal is that once your identity is tied to a wallet, all your transactions can be tied to you due to the transparent ledger. Since the introduction of Bitcoin, new currencies – such as Monero[1] – and new tools – such as crypto mixers – have been introduced with the aim of increasing the anonymity of trading cryptocurrencies. This increased anonymity is an advantage for people living in countries with heavy censorship, but this additionally means that criminals use cryptocurrencies for illegal activities.

2.2. Exchanges

Bitcoins can be mined or exchanged, but both require some technical knowledge. Cryptocurrency exchanges were introduced to make trading easier and to connect buyers and sellers. There are centralized exchanges (CEX) and decentralized exchanges (DEX). Both offer a hassle-free trading experience in exchange for transaction fees.

Centralized exchanges act as an authority, making sure that all transactions are legit. This offers some benefits to customers as these exchanges can offer customer support and recovery of funds when a trade goes wrong. Through these platforms, it is possible to exchange cryptocurrencies for fiat currencies, for example euros. Centralized exchanges must therefore adhere to Know Your Customer (KYC) procedures to combat money laundering. This means that users need to trust the CEX with their identity and that the success of the exchange platforms is dependent on their reputation.

Decentralized exchanges are fully autonomous. The central authority is replaced by smart contracts, which are open source transaction protocols. If a smart contract is created correctly, which can be verified due to it being open source, then they are a more secure transaction method than using centralized exchanges. However, some knowledge about smart contracts is required in order to be able to verify that the contract is actually doing what it claims to do. There is no way to recover funds that

[1] https://www.getmonero.org/

are wrongfully transferred using a smart contract. On decentralized exchanges, it is only possible to trade cryptocurrencies for cryptocurrencies on the same blockchain and it is not possible to exchange for fiat coins. Unlike centralized exchanges, decentralized exchanges do not require KYC information. Thus, users can maintain complete pseudonymity.

2.3. Cryptocurrency scam methods

This is a short overview of cryptocurrency scams that are relevant to this research. Making a definitive classification is difficult since there is both a lack of taxonomy of scams and since there is often overlap between different scam methods [4]. Therefore, literature studies and websites collecting self-reports often have their own classification systems [4]. Additionally, websites for self-reports have the uncertainty of victims wrongly interpreting what kind of scam they have fallen victim to. This makes quantification of cryptocurrency scams per category difficult. This study aims to adhere to the taxonomy created by Bartoletti et al. [4] and describes all scam techniques that are related to scams that make use of websites with an investment component.

Ponzi schemes are investment scams that claim to generate large returns. In reality, there is nothing generating these high returns. Thus, the platform can only pay investors their promised returns by getting new investors. The flow of new investments hides that there are no real profits made. These pyramid schemes can be very profitable to early investors, however once the new investments dry up, the scheme will inevitably collapse, leaving the more recent investors empty handed. A common way to continuously get new investments is to promise referral bonuses to current investors. More recently, these scams are being called High Yield Investment Programs (HYIPs). Cryptocurrency HYIPs can be offered through an investment website, or through smart contracts on Ethereum [4].

An example that illustrates the possible scale of a cryptocurrency Ponzi scheme is Bitconnect, which was active between 2016 and 2018. It created its own currency *Bitconnect coin*, which in 2017 was one of the top 10 most popular cryptocurrencies [12] , [13]. At its peak in 2017 the coin had a market capitalization of $2.5 billion. Once investors lost confidence, the coin lost all its value and dropped to less than a dollar per coin. Bitconnect's lead promoter in the United States pleaded guilty to conspiracy to commit wire fraud and was ordered to pay $17 million in restitution [14].

The interesting aspect of HYIP scams is that not every investor can be categorized as a scam victim, since there are people who willingly participate in the scheme even though they understand the mechanics behind them [15]. This is because they bet on being one of the early investors who do still get paid a very large return. There exist HYIP aggregator websites to aid such investors in figuring out if an investment website is paying returns or that they do not pay anyone at all.

Fake cryptocurrency services is a scam category that can be split into multiple services that are imitated. A prevalent scam is fake exchanges. These websites or apps look like actual exchanges, both CEXes and DEXes. Scammers can copy a well-known exchange or they can create a new exchange. In case scammers copy a well-known exchange, they often make use of typosquatting [16]. This means that the scammers copy a legitimate website and host it on a very similar looking domain name.

Multiple scamming methods can be tied to a fake exchange. An advance-fee scam convinces the victim to pay money upfront, in anticipation of receiving something of greater value. If a website promises a sign-up bonus that can only be withdrawn after paying fees such as transaction costs or capital gains taxes, then this could be part of an advanced-fee scam. Characteristic for an advanced-fee scam is that something of value is promised, but first requires upfront payment. The value of the promised return is much larger than the required payment, which makes it appear to be a good deal. In reality, the promised funds do not exist and the victim loses their investment, as well as the advance fees that were paid.

In many cases that promise great returns, victims see their investments grow over time. A dashboard on the website shows the current value of their investments and victims believe that this money is theirs to withdraw. However, the number that is supposed to represent their investments is not tied to any actual value. Scammers will use social engineering techniques to persuade the victim to invest several

more times. Once the victims want to withdraw from their account, they are shocked to discover that they cannot. In most cases, they are persuaded one last time to pay for transaction fees, taxes or some other made-up cost, before the scammers will break off all contact.

Another scamming method related to fake exchanges is phishing for KYC information. In this case, the fake exchange requires formal identification, similar to any regular CEX. The sign up information is used to build a profile of the victim. When a victim sends a copy of their identity documents, scammers can use this copy for other KYC verification processes or they can sell the identity information.

In order to gain new victims, pig butchering scammers can direct their victims to fake exchanges. Pig butchering scams, also known as romance scams or romance baiting scams, are characterized by scammers who invest into a personal relationship with their victim [10]. This can be both a romantic relation or an online friendship. The scammer chats with the victim to build trust, then uses this personal connection to deceive them into making cryptocurrency investments. In actuality, the victims are transferring money to the scammer. The scammer keeps extorting money until the victim understands that it is a scam or until the victim has no money left.

Lastly, a recovery scam is focused on victims of scams. Scammers try to convince victims who reported their victimhood online that they can help in recovering the lost funds. The scammers sometimes pose as a moderator of the report platform. If the victim accepts the offer, they must supply all their details, which can then be sold to other scammers. Since a recovery scam victim has been scammed or phished twice, this contact information is relatively valuable to other scammers.

2.4. Allocation of scam efforts

Scammers can create one website of high believability or they can cast a wide net by deploying many copies of the same website template, but compromising on believability. The two different approaches are discussed next.

In the case that scammers create a single website of high believability, it is more difficult to recognize the website as a scam. They can invest effort to create and maintain social media accounts and publishing supportive web articles to increase the authenticity of the platform. This increases the proportion of visitors that eventually fall for the scam. However, when there are many victims of one single website, then this also increases the chance that law enforcement starts an investigation. Additionally, if the website is taken offline, the full scam is unveiled. Creating a single website of high believability requires more customizations and effort than a more shallow website of lower believability. This strategy can therefore be characterized as *high risk, high reward*. In which *high risk* is relative to the overall low risk of prosecution.

A *low risk, low reward* strategy is creating many copies of the same website template. Because of the duplications, it does not require much additional research to figure out that it is a scam website. This decreases the proportion of visitors that fall for the scam. However, by spreading the victims across different websites, the police reports do not relate to the same website and are therefore less likely to incentivize an investigation. If a single website is taken offline, their operation on other websites can continue. Since the owners of these websites have invested less in their overall online presence, such as posting on social media platforms, their websites draw less overall traffic. This scamming strategy requires less attentive involvement, but is less likely to cause many victims.

2.5. The scam pipeline

In the aforementioned scams, the creation of the scam website itself is only one of the components in the scam pipeline. In general, a scamming operation faces challenges like any legal business does, but they do have some additional challenges. Like any business, in order to be successful, they need to draw customers, or in this case victims, to their websites. This can be done through advertisements, through social media contact, or through word of mouth. Additionally, they often offer customer service to answer questions or to help with transactions.

In some scams the victim pays with cryptocurrency immediately, in other scams, they leave their contact information in order to be called back. The victims are then persuaded to invest into cryptocurrencies

over the phone. After a victim has paid, the crypto funds needs to be laundered in order to hide their illegal origin.

A scam operation consists of different components: website creation, advertisement, customer support, and money laundering. These elements can be executed by the same entity or by different entities. In the second case, each entity gets paid for their specific services and receives either a fixed fee or a commission based on the success of the scam. This study focuses on the modus operandi of the owners of the website domain and of the creators of the website.

2.6. Shared responsibility

As mentioned in the introduction, law enforcement has the ability to start an investigation into a scam or to send a NTD. Anyone can send an NTD to a hoster, requesting the takedown of a website. If the hosting service provider does not comply, then the situation can be escalated to the registry. The registry is the organization behind a top level domain, such as the *.com* domain. In the case of *.nl* domains, this is the SIDN[2].

Apart from the NTD procedure, there are several other measures that can prevent someone from falling victim to a scam website. A browser, for example Firefox or Google Chrome, can place a scam warning screen in front of the website. This means that in order to continue to the website, visitors must click through the scam warning. Search engines try to optimize their algorithms in such a way to not suggest scam websites [17]. Content delivery networks (CDN) are services that aid in the distribution of websites to visitors efficiently. They too have the ability to display warning screens. Several anti-virus software programs use blacklists and display warnings when a user tries to visit a website labeled as a scam. However, accurate labeling remains difficult and therefore most of these prevention systems rely on user reports. Several websites, such as Chainabuse[3] and ScamWatcher[4] try to collect these scam domains. This collection relies on victim reports and on classification algorithms of active volunteers. Additionally, there are websites dedicated to explaining the workings behind scam operations and to warning people about the latest scams, an example of this is ScamAdviser[5]. Furthermore, an involved social circle can prevent someone from falling victim to a scam or to help them realize they are getting scammed [18].

This mix of possible measures to prevent scamming means that anyone can aid in making the internet a safer place, The downside of these various preventative measures is that nobody has the sole responsibility of creating a safer internet. In the end, the victim can hold nobody accountable other than the scammers.

[2] Stichting Internet Domeinregistratie Nederland: https://www.sidn.nl/en
[3] https://chainabuse.com
[4] https://www.scamwatcher.com/
[5] https://www.scamadviser.com/

3

Related works

In the field of cryptocurrency scam research, the following key contributions have been made.

In 2012, three years after the introduction of Bitcoin, Moore et al. [15] described the concept of HYIP scams. They analyzed the truthfulness of HYIP aggregators and provided policymakers with advice on how to intervene.

Drew and Moore then focused on clustering criminal websites [19], [20]. Since criminals want to scale their scams, they create duplicate websites with only minor differences. Finding such clusters can allow for shutting down scams more quickly and "[help] tackle cybercrimes that individually are too minor to investigate but collectively may cross a threshold of significance" [19, p. 12]

The 2017 paper written by Toyoda et al. [21] proposed a technique to classify Bitcoin addresses as fraudulent, based on previous transaction patterns. They concluded that the frequency of transactions is a very effective feature for the classification of Bitcoin addresses used for HYIP scams. In the 2018 paper written by Toyoda et al. [22], the classification of HYIP scams is improved by incorporating a Bitcoin price volatility effect mitigation. They compared five classifiers and reached a performance of 95% true positives, while only having 4.9% false positives. The drawback of identifying HYIP scams based on previous transaction patterns of Bitcoin addresses is that there will always be victims before problematic addresses can be identified.

Chen et al. [23] analyzed the working of Ponzi schemes in Ethereum smart contracts. They created a classifier to detect intentionally obfuscated Ponzi schemes in Ethereum smart contracts. The classifier is based on the operation codes of a smart contract and can flag malicious contracts before they are executed.

Siu et al. [24] have performed a study that focused on advertisements for scams on the platform Bitcointalk. They identified a peak in HYIP advertisements in 2018 and again in 2021. They observed a shift in popularity from the keyword "Ponzi" to the keyword "HYIP".

Zhang and Yan [25] built a classifier for detecting malicious websites by creating a neural network for analyzing the URL of websites. Although this study focused on malicious websites in general, this classifier can be trained on cryptocurrency websites specifically. This language model is not yet able to analyze the contents of websites. Since a suspicious URL is on itself not enough evidence of wrongdoing, the model should not be used as only substantiation for sending NTDs.

A concept that was frequently used during this study was the concept of deviant security practices, introduced by Van de Sandt [26]. This concept described the way that cyber criminals protect their business operation against their adversaries, among which are law enforcement, competitors and internet vigilantes. By taking on the perspective of the cyber criminals, one can reason about their ways to protect assets and to remain anonymous. Van de Sandt [26] analyzed the security of criminal operations based on concepts such as the CIA triad of confidentiality, integrity, and availability, which is originally used to model the security challenges that legitimate businesses face.

4

Research gap

Previous research focused on describing scam phenomena and improving scam website classifications. These classifiers all had their own limitations. They were either only useful for post-hoc classification of scams, focused on the URL only – which is not a sufficient basis for sending an NTD –, or were not tied to a website, but to other scamming means, such as smart contracts. Commercial classifiers exist, but are not transparent in their feature extractions and therefore not a sufficient basis for sending NTD's.

This research will therefore focus on discovering and describing the Tactics, Techniques, and Procedures (TTP) of the owners of scam websites. By describing their TTP, this research provides knowledge that helps with the early and automatic detection of scam websites. Early detection should help reduce the number of victims of a scam website and automatic detection reduces the time and effort needed from law enforcement to assess whether a website should be taken offline. In addition, this research will aim to address the issue surrounding the feasibility of criminal investigations by trying to capture which aspects can predict the potential for a successful prosecution.

The overarching question that this study aims to answer is:

What can be learned from the analysis of scam website clusters about the Tactics, Techniques, and Procedures of the owners and creators of cryptocurrency investment scam websites?

This study aims to answer this question supported by the findings of the following sub questions:

What does the life cycle of a crypto investment scam look like?
What can be learned about the TTP from analyzing website clusters?
What tools and services do scammers use to create their websites?
What tools and services do scammers use to obfuscate their identity?

5

Measurement Approach

This research focuses on extracting knowledge about the TTP of the owners and creators of crypto scam websites. The content of the websites, as well as the metadata of the websites are analyzed to extract information about the lifespan of websites, as well as to provide an understanding of the similarities between websites.

Due to the limited lifespan of scam domains, previously constructed datasets become outdated quickly. Kaggle offers datasets intended for machine learning purposes that contain the domain names of scam websites, but many of these websites are already offline [27]. This makes the dataset useful for training domain name based machine learning models, but not for analyzing the contents of the websites. Xia et al. [16] created a cryptocurrency exchange scam dataset by leveraging typosquatting, but many of these domains are also not online anymore. A recent dataset that contained HTML data of crypto scam websites did not seem to be available. Therefore, it is a necessary step to construct a new dataset of domain names in order be able to analyze the contents of these websites.

5.1. Constructing domain name dataset

In order to construct the dataset, several websites that collect scam reports have been reviewed, among which are ScamAdvisor and Chainabuse. This provided valuable insights into various recent or ongoing scams. While many of the scam websites were unique, ScamAdvisor listed multiple websites that all looked very similar. The websites looked sophisticated and were on itself difficult to be identified as a scam. However, seeing all websites together, it was very clearly part of a scam. The clearest indication was that the alleged founder of these websites were all the same person, yet there was no digital trace of him anywhere on the internet.

This finding sparked the question whether these scam websites were owned by one actor or if the template was part of a scamming kit that could be bought by anyone. Since many of these websites are still online and new scam domains are still being created, this research investigated the clusters of similar cryptocurrency scam websites.

By looking at the website and predicting which lines would probably not appear on other websites, possibly distinct phrases were selected. A Google Search query with the possibly distinct phrase, within quotation marks in order to only suggest exact matches, either confirmed or contradicted these hypotheses. If the query suggested websites that had a different content, it was deemed a non-distinct phrase. If the phrase seemed to only appear on websites that shared the same content with the initial website, then it was deemed a distinct phrase. To clarify, this can identify websites from the same template as the initial website or it can lead to clusters of websites that have a different template, but share the same content.
In the case that a possibly distinct phrase only leads to the initial website and to no other websites, then

it was a distinct phrase, but not of any use to discovering more scam websites. An example of this is phrases that contain the domain name of the website.

5.2. Template descriptions

Using this method, four templates that share at least some of the content are identified. The identified distinct phrases form the seed queries to collect the dataset. All four templates use a Cloudflare anti-bot front page to limit automatic access to their websites. The implications of this are discussed later. The websites use the advanced-fee method to scam money and can ask for KYC information. In the next section, the four templates that are analyzed in this study will be discussed.

5.2.1. Template 1

This first template occurs the most often. An instance of this template is shown in figure 5.1. It contains a slider with text and illustrations on top of the website, the current market trends in the middle, and more explanation on how the registering and investment process works at the bottom of the page. The landing page of template 1 is the simplest of the four.

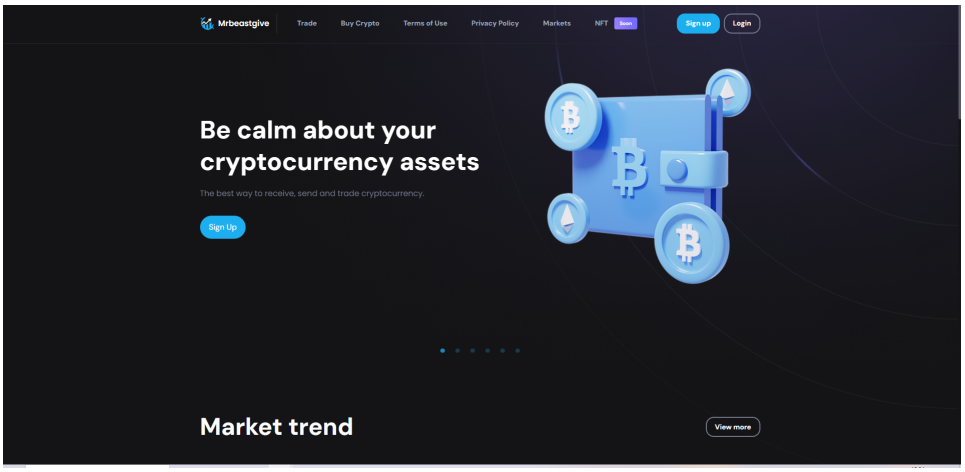


Figure 5.1: The landing pages of template 1 as of 7/10/2024.

The distinct phrases that can be found in this template are:
Phrase 1: “Crypto trading with a leading exchange”
This phrase is chosen because if the exchange were indeed leading, then it is not likely to appear on many websites.
Phrase 2: “A team of specialists around the clock will help and answer any question”
This phrase contains a grammatical error. It is worded in such a specific way, that it is not likely to appear on other websites.
Phrase 3: “Be calm about your cryptocurrency assets. The best way to receive, send and trade cryptocurrency.”
This phrase is using “Be calm” to provide reassurance, which is common for scams. The two sentences together form a distinct phrase.

Phrases that did not work for this template were “Hank Weizen” or variations of the sentence “Recognizing the importance of Bitcoin from the onset, and understanding that the exchange is the most critical part of the cryptocurrency ecosystem, Hank Weizen founded [<website name>] ...”. The <website name> varies depending on the website. Although the name Hank Weizen is unique to these scam websites, it appears in too many articles that report the scam, since it is the item that signifies the untrustworthiness of the website. The full sentence in which this name appears is taken from the cryptocurrency trading platform Kraken[1] and can therefore also not be used to confirm scams.

[1] https://www.kraken.com/

5.2.2. Template 2

The second template is discovered because it shares all three distinct phrases. It has a similar slider as template 1, but uses different illustrations and neon green as the main color. Below the market trends it contains a list of services that should convince visitors to start using the platform, a guide through the registration process, and claims about the popularity of the platform. An example of template 2 is shown in figure 5.2

Figure 5.2: The landing pages of template 2 as of 8/10/2024.

5.2.3. Template 3

Template 3 was discovered because it shares the third distinct phrase. The website has different subpages and seems to focus on NFT trading. The landing page of the website is more densely packed with information than that of template 1 and 2. They do not have a slider at the top of the page like template 1 and 2, but have some references to legitimate cryptocurrency exchanges. Below, they show the market trends, aim to convince to visitor to start investing and make claims about the popularity of the platform. They seem to target the more experienced traders and promote with having an API and accurate anti-money laundering checks. They have a tool to calculate expected revenue over time, explain their characteristic features, and end with a cryptocurrency conversion tool. Figure 5.3 shows an example of a website with this template.

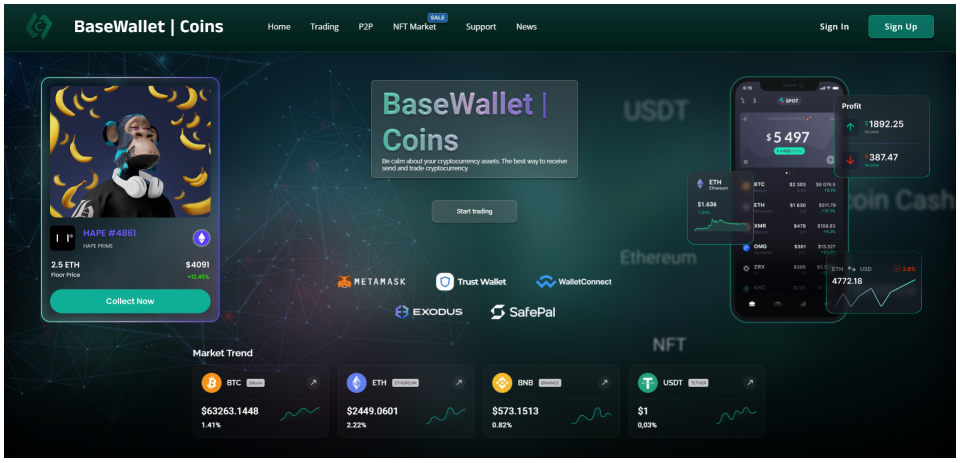


Figure 5.3: The landing pages of template 3 as of 8/10/2024.

Since this webpage contains some different content, distinct phrases 1 and 2 could not be linked to this template. However, an additional phrase that identified websites with this template was found:

Phrase 4: “The Transactions offered by this Website can be used only by fully competent adults. Transactions with financial instruments offered on the Website involve substantial risk and trading may be very risky”

This line was selected because the phrasing “only by fully competent adults” is notable, as well as “involve substantial risk and trading may be very risky,” which is stating that risk is involved twice.

5.2.4. Template 4

The fourth template shares the first distinct phrase. The landing page has a toggle to view it in light mode or dark mode. It uses a purple and green layout and shows a clean illustration with the title and a sign up button. Below, it shows the current cryptocurrency values, followed by some special and temporary offers, a three-step guide to earn money, and the current market trends. Lastly, it shows their security efforts, the popularity of the platform, some unique selling points and finishes with a registration button. Template 4 is shown in figure 5.4.

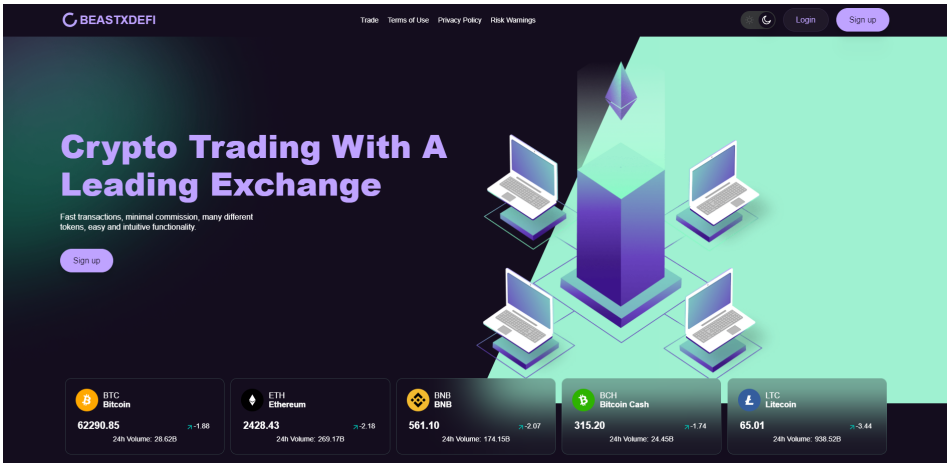


Figure 5.4: The landing pages of template 4 as of 8/10/2024.

This template offers limited possibilities to find distinct phrases due to the content being shared across multiple different scam websites that did not necessarily adhere to any template. Therefore only the first distinct phrase could be linked to this template.

5.3. Search engine comparison

Although using a search engine has some negative consequences for the transparency of the creation of the dataset, there was no suitable alternative. Web crawling needs a seed and leverages links on the seed webpage in order to get to new web pages and websites. If a website is not referenced on any other website, then it is generally not found. Given that search engines specialize in web crawling and have developed methods to find such ‘hidden’ websites, their indexing was used in order to expand the dataset. Websites can prevent search engines from indexing their websites, but since these scam websites generally want to be found by potential victims, the assumption was made that they do not actively block being indexed by search engines.

Seven search engines are compared in order to establish which search engine could be best used to discover new websites with any of the four templates as described above. The seven engines that were evaluated are: Bing, Brave, DuckDuckGo, Google Search, Qwant, Startpage, and Yandex. The assessment focused on two critical factors: the availability of an API and the ratio of accurate search results. The search engines were testing with the search term *“crypto trading with a leading exchange”* with quotation marks. Table 5.1 shows the number of results and whether the search engine has an API.

Table 5.1: Search engine comparison

Search engine	Number of results	Availability of API
Bing	28, of which 9 refer to relevant sources	Yes
Brave	0	Yes
DuckDuckGo	11	Yes
Google Search	approximately 9000*	Yes
Qwant	26, of which 8 refer to relevant sources	No
Startpage	100	No
Yandex	1	Yes

DuckDuckGo offers an API, but it is unclear whether this is maintained by the same organization. Google Search claims to show approximately 9000 results, but this number cannot be verified. They offer an API that can provide 100 results per search query for free. The paid version of the Google Search API allows for more queries per day, but does not offer more results per query. Further investigation into Startpage confirmed that their search engine makes use of the Google Search API. Yandex shows 1 result and has an API. However, due to their sensitive geolocation, they were disregarded in this decision.

Overall, Google Search seems to offer the best quality and quantity of results and offers a partially free API. Therefore, the Google Search API is the most suitable choice for this use case. However, it is still a black box model and how exactly Google does their indexing is a well-kept secret. Google Search does aim to prevent suggesting spam websites [17]. Thus, some domains are possibly removed from the suggestions by Google. It is difficult to determine what the effect of this is on the representativeness of the dataset for the whole scam operation.

5.4. Domain name collection

Through the Google API, the top 100 domains for a Google Search query are collected. Because of the API limitation of 100 domains per query, this is done by querying the four different distinct phrases within quotation marks and over several days during a three month period between June and August. Although the key phrases are carefully chosen, the Google API sometimes suggests legitimate websites. However, this was overall less than 10 websites and the labeling step filtered out these domains. Additionally, 20 manually found domains are added to the list. These are found through scam report warnings or through advertisements on social media. This results in a list of 513 domains that likely belong to the four templates.

These 513 domains are manually labeled based on the contents of the website and categorized on whether they belong to one of four templates. An overview of the labeling is shown in table 5.2. The 436 domains that belong to templates 1 to 4 form the basis of the dataset. These domains can be seen in A.

Table 5.2: Overview of the website labeling

Label	Number of Websites
Template 1	282
Template 2	73
Template 3	58
Template 4	23
Different scam website	29
No scam website or offline	48

5.5. Metadata collection

The created domain dataset is then used to collect more information about these websites. Apart from the contents, which is discussed later, the metadata is collected. The main goal is to collect information

about whether the websites are still online. The metadata is collected with a Python script that is manually run daily around 12 am (GMT+2) between 02/09/2024 and 11/10/2024. A commercial VPN with Dutch geolocation is used, which was tested to confirm that the extra latency did not impact the quality of the results. The VPN is used to hide and protect the researcher's own IP address. The Dutch geolocation was chosen since this study is done in collaboration with Dutch law enforcement and would best mimic the internet access of Dutch citizens.

The script, which is based on the code of a previous project of the researcher with modifications to suit this research, collected and verified the website certificate data, the IP address with its geolocation and received status codes. This is done for *https://domain.tld* as well as for *https://www.domain.tld*, in which *domain* and *.tld* are placeholders for the actual domain names and top level domains, such as *.com*. At the start the data gathering (02/09/2024) an additional 6 websites were offline. Resulting in 430 website of which at least one snapshot of metadata was recorded.

Domain registrar and registrant data are collected at the beginning of the measurements, but not daily as the information remained relatively static. The Whois tool[2] is used to receive this data. Due to the non-standardized format, this data is parsed to extract the relevant information to JSON format. A disadvantage of this method is that it does not work well for all top level domains, but since the vast majority of the dataset is *.com*, this is sufficient. Alternatively, existing domain tooling can be used for this purpose, which work for all top level domains and have already applied the parsing steps, but these tools are either paid services or it is not possible to request for multiple domains at once.

5.6. Content collection methods

To collect the content of the website in order to create a HTML dataset, a web scraper is used on several webpages of each of the websites. The main argument for scraping the contents over utilizing internet archiving initiatives such as the Wayback Machine[3], archive.is[4] or Common Crawl[5] is that these archiving tools tend to only capture the Cloudflare front and not the website contents behind the Cloudflare front.

5.7. Cloudflare Turnstile

The Cloudflare anti-bot page is called a Turnstile page [28, 29]. This is a free tool that is used to tell bots and humans apart and only grants access to the website if the entity trying to gain access is human, thus preventing automated access and DDoS attacks. Unlike a CAPTCHA, no puzzle needs to be solved to prove oneself human, which allows humans to more rapidly surf to websites. It is more inclusive for visually impaired people and overall reduces frustration.

The Turnstile tool runs "a series of in-browser tests, checking browser characteristics, native browser APIs, and asking the browser to pass lightweight tests [...] to prove that it's an actual browser" [28]. However, the precise implementation remains a proprietary secret.

Initially, the scraping was done by using Selenium[6] with Python. Using this method, it was unexpectedly possible to get through the Cloudflare front and scrape screenshots and content. However, at that point in time the domain dataset was not yet finalized. At some point between June 5th and June 11th 2024, this method stopped working, which is likely due to a change in the verification method. Since the verification procedure is secret, it was not possible to deduce what the change was and how to overcome this. Multiple different approaches were attempted, among which were accessing the website through command line and various implementations using the Python packages Selenium-stealth[7]

[2] https://learn.microsoft.com/en-us/sysinternals/
[3] https://web.archive.org/
[4] https://archive.is/
[5] https://commoncrawl.org/
[6] https://pypi.org/project/selenium/
[7] https://pypi.org/project/selenium-stealth/

and Flaresolverr[8], or the JavaScript library Puppeteer[9], but none were able to bypass Cloudflare.

An option that did end up working was using PyAutoGUI[10] – which is used to automatically perform mouse and keyboard actions –, with the general Google Chrome browser, which had a well-established user profile due to long-time usage. However, the risk of using one's own browser for scraping is it can lead to permanent blockage, which can be troublesome. Luckily, it turned out that it was not the user profile that made the turnstile test succeed, but the legitimacy of the browser. Therefore, PyAutoGUI in combination with a clean Firefox browser without any user history succeeds into getting through the Cloudflare Turnstile page. Firefox is selected due to its built-in screenshot feature. However other browsers should work similarly.

The new implementation opens a Firefox browser and navigates to a given website. PyAutoGUI lets the cursor click at the right locations in order to take a screenshot and to save the HTML contents from the Inspect pane. Using this approach, the dynamically loaded contents are captured in the HTML contents. Only the pages with textual content were scraped. Pages that exclusively displayed an API were ignored to optimize scraping time. A simple opcode structure is defined and an opcode instruction file per template is created to specify all clicks, scrolls, waits, and file-save instructions. The coordinates in the opcode instruction files are system-specific and may need to be altered when implementing on a different device.

Depending on the template, if an unexpected template difference occurred, the scraper would either terminate, which is favorable, or continue, thus saving the wrong contents. The scraper therefore needed manual oversight to a mild extent in order to address terminations for template 1 and 2 and to verify the correct behavior for template 3 and 4. Still, this automated method was quicker than manual data collection and spent approximately one and a half minutes to process a full website.

The eventual scraping was performed between 02/09/2024 and 04/09/2024. 34 websites that were online on 02/09/2024 during the metadata collection were already offline when their content was scraped. Therefore, a dataset of 396 websites is used for the similarity analysis. Since the initial labeling confirmed the template, these 34 websites are not entirely removed from the dataset and can be used for metadata analysis. The number of websites per template used for the similarity analysis is shown in table 5.3.

Table 5.3: Number of websites scraped per template

Label	Dataset Size
Template 1	276
Template 2	66
Template 3	31
Template 4	23

5.8. Legality of scraping

Scraping itself is not a illegal activity by definition, but it is a legislative gray area in which caution is needed. In general, web scrapers should adhere to all GDPR regulations, copyright laws and to the terms of service of a website.
Generally, websites include Robot Exclusion Protocols at *https://domain.tld/robots.txt*, which dictates what pages can and cannot be scraped. Although it might be up for discussion whether the *robots.txt* of malevolent websites should be adhered to, this scraping did comply with the Robot Exclusion Protocols of the websites that had such a *robots.txt* page. None of the content was behind a login and the pages did not contain any personal data. The scraping was thus conducted with a strong belief that it remained within all legal boundaries. The scraped data is solely used for this research and is not publicly available. Although this research is done in collaboration with law enforcement, the contents of the scrapes were not used for any criminal investigations.

[8] https://github.com/FlareSolverr/FlareSolverr
[9] https://github.com/puppeteer/puppeteer
[10] https://github.com/asweigart/pyautogui

6

Methodology

The collected data can now be used to extract characteristics of these scam websites and to analyze the similarities and differences within and between templates. This section first explains the metadata analysis, after which the steps taken to process the content data are described.

6.1. Mapping the lifespan

In order to calculate the average lifespan of the websites, it is necessary to be able to establish when a website goes offline. The first attempt used the HTTPS availability of a website as metric for websites going offline. However, this metric did not work. This is because websites that go offline – either due to being taken offline or due to a domain or hosting contract expiring – often still display a notice informing visitors that the website is currently offline.

A second approach did work. Most websites in the dataset use a Cloudflare Turnstile page, which operates as a proxy server. It is made to handle large quantities of web traffic and to only let the traffic of legitimate visitors through to the actual website. Therefore, the IP address of the actual website is only known to Cloudflare and visitors are shown a Cloudflare owned IP address instead. If this Cloudflare service is discontinued due to the website being taken offline, then the IP address shown to visitors is no longer a Cloudflare owned IP address. Thus, by looking at ownership of the IP address, it is possible to register a website going offline fairly accurately.

The predictions made using the Cloudflare IP address as metric are compared to the manual assessment made on October 11th 2024. This comparison verified that 99.07% of domains were correctly classified as online or offline. Which means that availability of a Cloudflare IP address is an accurate way of establishing the event of a website going offline. Table 6.1 shows the confusion matrix describing this comparison. The two websites that are predicted to be offline, but observed as online, did not use Cloudflare services. The two websites that are predicted to be online, but were offline, did use Cloudflare, but have been repurposed and deliver different contents.

Table 6.1: Confusion matrix for Cloudflare IP prediction

Prediction / Observation	**Online**	**Offline**
Online	311	2
Offline	2	115

The lifespan of the websites is calculated by taking all websites that have gone offline between the first and last metadata measurement and computing the difference between their creation date and the last date that the website used a Cloudflare IP. Furthermore, the total days online is taken for all websites that are still online in order to verify that the lifespan for the websites that went offline is representative for the dataset as a whole.

6.2. Content data analysis

The website scrapes are sorted into four templates, based on the manual labeling as described in the measurement approach. The website similarities are calculated within templates and between templates, respectively named the *intra-template similarities* and the *inter-template similarities*. Two metrics are used for the comparison of any two webpages: a structural similarity and a textual content similarity. This split is illustrated in figure 6.1. Since these metrics are used in both the intra-template and inter-template similarity analysis, these are described first. Afterwards, the intra-template and the inter-template similarities are explained.

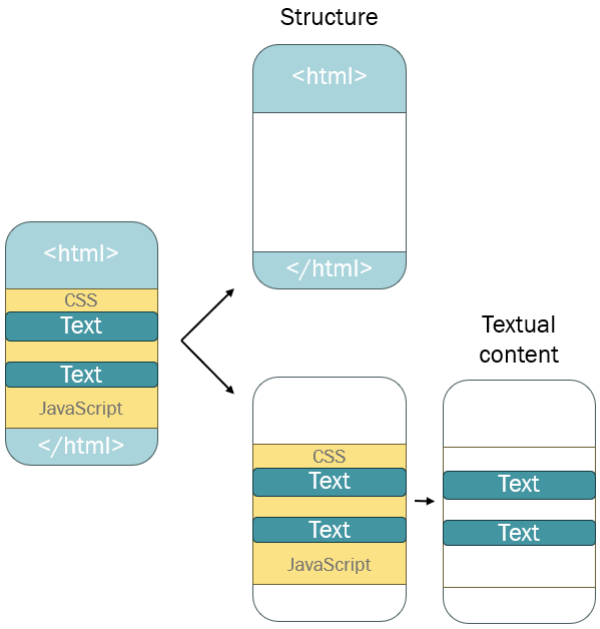


Figure 6.1: The split between applied similarity metrics visualized.

6.3. Structural similarity

The structure is defined to be all HTML tags on a webpage. HTML tags generally define the structure of a website and how this is visually represented in the browser. In order to calculate the structural similarity of two webpages, the *html-similarity* Python package[1] was used. The package contains a *structural_similarity* function, which is based on a 2016 paper by Gowda and Mattmann [30]. This paper proposed a way to calculate structural similarities based on the Tree Edit Distance between two Document Object Model (DOM) trees. In order to reduce the computational intensity, the html-similarity package calculates the similarity between two HTML tag sequences instead. The sequences evaluated are the pre-order traversals of the HTML DOM tree. The function then calculates the similarity of these sequences using Python's built-in difflib SequenceMatcher which is based on Gestalt Pattern Matching, also called Ratcliff/Obershelp pattern recognition [31].

Gestalt pattern matching aims to match the largest common part of two sequences. Afterwards, there are two recursive calls on the sequences left and right of the largest common part. This is continued

[1] https://pypi.org/project/html-similarity/

until there are no more common parts. The similarity S between a and b is then defined by taking the number of common parts, K, multiplying this by 2 and dividing it by the sum of lengths of a and b. This results in a number between 0, for no similarity, and 1, for identical sequences. The definition is shown in equation 6.1.

$$S(a,b) = \frac{2 \cdot K}{|a| + |b|} \tag{6.1}$$

If there are multiple largest common parts, the *SequenceMatcher* function chooses the leftmost largest part of a as largest common part. This has as a consequence that computing $S(a,b)$ can have a different outcome than $S(b,a)$. In order to ensure that the calculated similarity is commutative, the similarity is taken as $MAX(S(a,b), S(b,a))$.

The advantage of this structural similarity calculation is that it takes the order of tags into account. A set based calculation, such as the Jaccard similarity which is defined as the size of the intersection divided by the size of the union, does not take any ordering into account. Since there is not much variety in tags expected, the information regarding the order of the tags matters and should be reflected in the similarity score.

6.4. Textual content similarity

For content evaluation, the webpage is mapped to Term Frequency - Inverse Document Frequency (IF-IDF) vectors. The cosine similarity of two webpages are calculated by taking the dot product of the normalized TF-IDF vectors. This is a commonly used metric for comparing longer text bodies. If two non-empty documents are compared, then the metric returns a value between 0 and 1, where 0 means that there is no similarity at all between the webpages and value 1 means that they contain identical content. The order of the content does not influence the similarity score.

At first, an approach in which everything that was not in HTML tags was taken into account. This would yield all the text on the webpage, as well as any CSS and JavaScript on the page. However, it was noticed that the relatively large quantity of CSS and JavaScript, compared to textual content on webpages, made comparisons of texts difficult. Since the CSS and JavaScript were the same across a template, the similarity scores were inflated to around 99%. Removing CSS and JavaScript from the comparison allowed for a more informative analysis. This is done by stripping HTML tag-free text of all lines that end in a semicolon or contain a curly bracket '{' or '}' and of all lines that start with an exclamation mark ('!'), at sign ('@'), period ('.'), forward slash asterisk ('/*') or double forward slash ('//'). This removed most of the scripting and left most of the text.

In order to verify this method, a comparison is performed for each of the templates. For each of the templates, a representative sample website is evaluated. These are the same websites as used in the inter-template analysis. An explanation of how these websites are chosen is provided in section 6.7 and section 7.4. These also explain why the verification of template 2 is split into parts a and b. The verification is done by manually stripping the webpages of the representative sample and comparing this to the automatically stripped webpages. The cosine similarity of the automatically stripped and manually stripped webpages are calculated and averaged. The averages for each of the four templates is shown in table 6.2 and should approach 1. The content of template 4 was constructed in such a way that it was more difficult to separate code from text, therefore this average similarity is lower than that of the other templates.

Table 6.2: Verification of automatic stripping method

Template	Manual/Auto cosine similarity
1	0.998
2a	0.995
2b	0.996
3	0.964
4	0.884

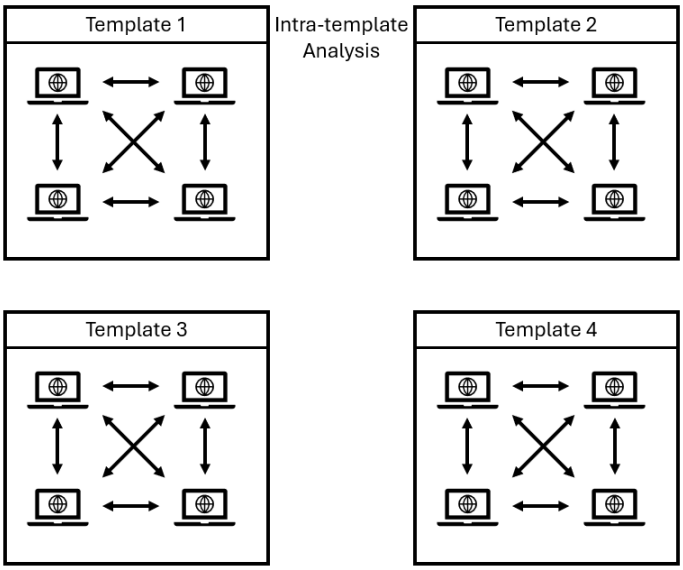


Figure 6.2: A schematic of the intra-template analysis.

6.5. Intra-template analysis

Now that the two similarity metrics are known, the intra-template analysis discussed. The intra-template analysis looks at the similarities between websites within the same template. A schematic of this is shown in figure 6.2.

For each website, there are between seven and ten webpage scrapes: one for the landing page and one for each of the subpages that mainly contain text. Since the subpage labels are the same within a template, e.g. *about-us* or *FAQ*, the pages can be compared to each other.

In one case, direct comparison of pages is not possible because the *buy-crypto* page in templates 1 and 2 is not present on all websites. When a webpage for a particular website is missing, its similarity with the corresponding pages of other websites is considered undefined.

For each template and page label, all pairwise similarities between different websites within a template are calculated for both structural and textual content similarity. There are 10 page labels for both template 1 and 2, 7 for template 3 and 9 for template 4. Thus, 36 page labels in total. There are 2 similarity metrics applied per page label, which results in a total of 72 tables of pairwise similarities.

The information of these 72 tables is condensed by taking the mean, standard deviation and median of each of the tables. Since the pairwise similarities are commutative, these statistical values can be calculated based on the values in the upper triangle of a table alone. The diagonal is excluded in this calculation, since it contains the value of a webpage's comparisons to itself, which, by definition, will be 1. The statistical values are then analyzed in order to observe relations and anomalies.

Additionally, the pairwise similarities of the full websites are calculated. These are defined to be the average of all pairwise similarities between two websites. If a pairwise similarity is undefined due to a webpage not existing, then this is either counted as a similarity of 0, whenever the webpage did exist for one of the two websites, or completely disregarded, whenever the webpage did not exist for either website. This resulted in 8 tables, which are condensed by taking the mean, standard deviation and median of each of the tables, following the same method as described above.

6.5.1. Influence of domain name

It is hypothesized that the only difference between websites of the same template is the domain name. Therefore, it is interesting to investigate the influence of the domain name on the intra-template content similarity. This can be done in two different ways. In the first, the HTML processing step is extended by taking the domain name of the website and replacing all occurrences of the domain name in the HTML file with a placeholder that is equal for all websites. This processing step causes the content similarity values to be equal or higher than the original content similarity values since the different domain names are now represented by an equal placeholder. Since it is expected that the domain name appears at least a couple of times on each of the webpages, this might depict a unrepresentative similarity value. In the second method, the HTML processing step is extended by removing all occurrences of the domain name. This method thus takes out the influence of the domain name on the similarity score entirely. When one component of a similarity score is removed, the influence of the remaining components scales up because they now represent a larger proportion of the total score. However, this method does not capture a difference in domain name occurrences between websites.

The occurrences of the domain name are identified by considering the domain name in all-uppercase, with the first letter uppercase, and in all-lowercase. If the domain name is on the website represented by multiple words, then this is not replaced or removed.

6.6. Template cores

In the inter-template similarity analysis, the similarities between templates are analyzed. A schematic of this can be found in figure 6.5. In order to perform this analysis the templates are represented by a single sample website. To ensure that the sample is representative for the template, clustering techniques are employed to find the templates core cluster. This way, the sample represents the template during the inter-template analysis as accurately as possible.

The pairwise average intra-template similarities of each of the templates are used in order to create an idea of which websites belong to the core of the template cluster and which are more loosely connected to the template cluster. The clustering happens based on a minimum similarity threshold. For a given template, similarity metric (structural/textual content), and threshold, all clusters are computed and displayed in which the mutual similarities are above the threshold. The chosen step size of the threshold values were not uniform, as adjustments are made in an ad-hoc manner. Therefore, this process is recorded in detail in the result section.

If there is a clear disjoint distinction between a large cluster and one or more smaller clusters, then the large cluster is more representative for this template than the small cluster. In this variation on divisive clustering, the smaller cluster is eliminated from taking part in the next iterative step. By iteratively taking slightly larger thresholds and by alternating between structural similarity and textual content similarity, the core of the template could be established. In the example shown in figure 6.3, it can clearly be seen that the cluster of website ID's $\{0, 75, 82, 95, 257\}$ are the outliers that will be eliminated from taking part in the next iterative step.

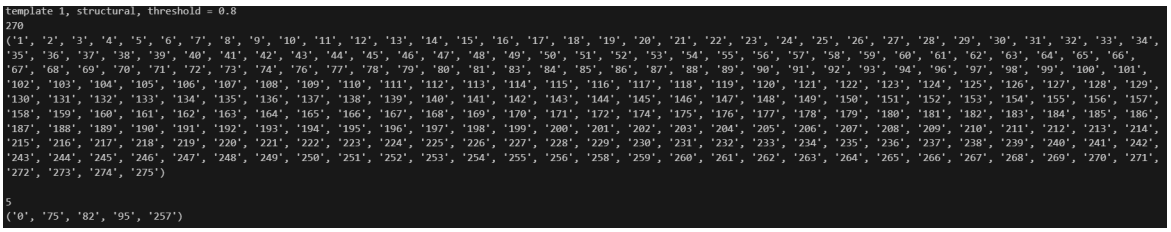


Figure 6.3: Example output of a division-based clustering step.

Alternatively, if there were no disjoint clusters, but multiple larger clusters existed, then the intersection of all those clusters was taken to construct the core cluster and inverse of this intersection was elim-

inated from taking part in the next iterative step. Again, the final core cluster is found by alternating between clustering based on the structural similarity scores and based on the textual content similarity scores. An example of this method can be seen in figure 6.4. The eliminated elements of this step are $\{0, 9, 12, 23\}$.

```
template 3, structural, threshold = 0.8
31
('0', '1', '2', '3', '4', '5', '6', '7', '8', '9', '10', '11', '12', '13', '14', '15', '16', '17', '18', '19', '20', '21', '22', '23', '24', '25', '26', '27',
'28', '29', '30')

29
('0', '1', '2', '3', '4', '5', '6', '7', '8', '9', '10', '11', '13', '14', '15', '16', '17', '18', '19', '20', '21', '22', '24', '25', '26', '27', '28', '29',
'30')

29
('1', '2', '3', '4', '5', '6', '7', '8', '10', '11', '12', '13', '14', '15', '16', '17', '18', '19', '20', '21', '22', '23', '24', '25', '26', '27', '28', '29',
'30')
```

Figure 6.4: Example output of an intersection-based clustering step.

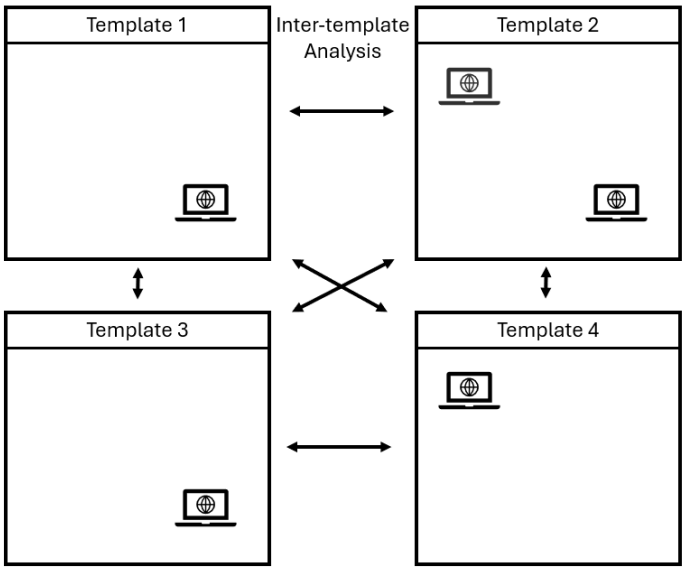


Figure 6.5: A schematic of the inter-template analysis.

6.7. Inter-template similarity analysis

From the template cores, one sample website is randomly selected to represent the template. Because of the clustering step, this is a representative sample. These websites are analyzed together in terms of structural similarity and content similarity. Since the labels do not fully overlap across templates, a pairwise comparison is performed between all webpages from all templates, regardless of the label and which template it belonged to. The same similarity metrics as for the intra-template similarity analysis are used. Thus, the result of this step is two tables with pairwise inter-template similarities: one for the structural similarities and one for the textual content similarities. The tables are then analyzed to extract patterns and anomalies.

7 Results

This chapter presents the results from the metadata analysis and the content data analysis.

7.1. Lifespan

The domain creation dates are extracted from the website's domain registration information. The heatmap shown in figure 7.1 shows the intensity of domain creations per month per template. Note that only the months in which at least one domain registration has taken place are displayed on the x-axis. The percentages depicted are the intensity of domain registrations per template. Thus, the percentages of one row add up to 100%. Template 1 has a high domain registration intensity between September 2023 and December 2023. Domains of template 2 in this dataset are mainly created between September 2023 and November 2023. Trends are less obvious for templates 3 and 4. Across templates there is a lower domain intensity in the time frame up to September 2023.

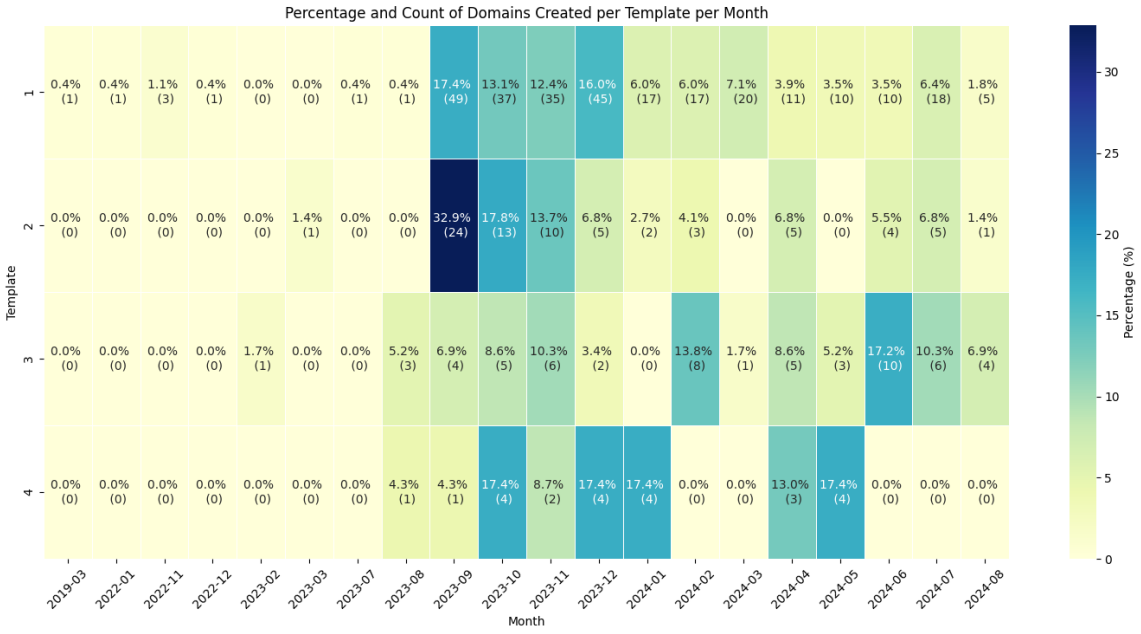


Figure 7.1: Percentage and count of domains created per template per month as of 02/09/2024.

Figure 7.2 depicts the domains that went offline between September 2nd 2024 and October 11th 2024. Going offline is defined as not showing the template anymore. In most cases, this means that the website does not show any content, but in 2 cases it showed different content. The figure shows the number of domains that have gone offline and the percentage of the total websites of the template that were created in that month. The coloring is based on the percentage, thus, if a template-month combination does not exist, then no percentage can be computed and it will be a white cell. This coloring shows very clearly that the majority of the websites that were created exactly a year ago went offline. Additionally, it shows that domains of template 3 are being taken offline more frequently, in terms of percentages, than the other templates.

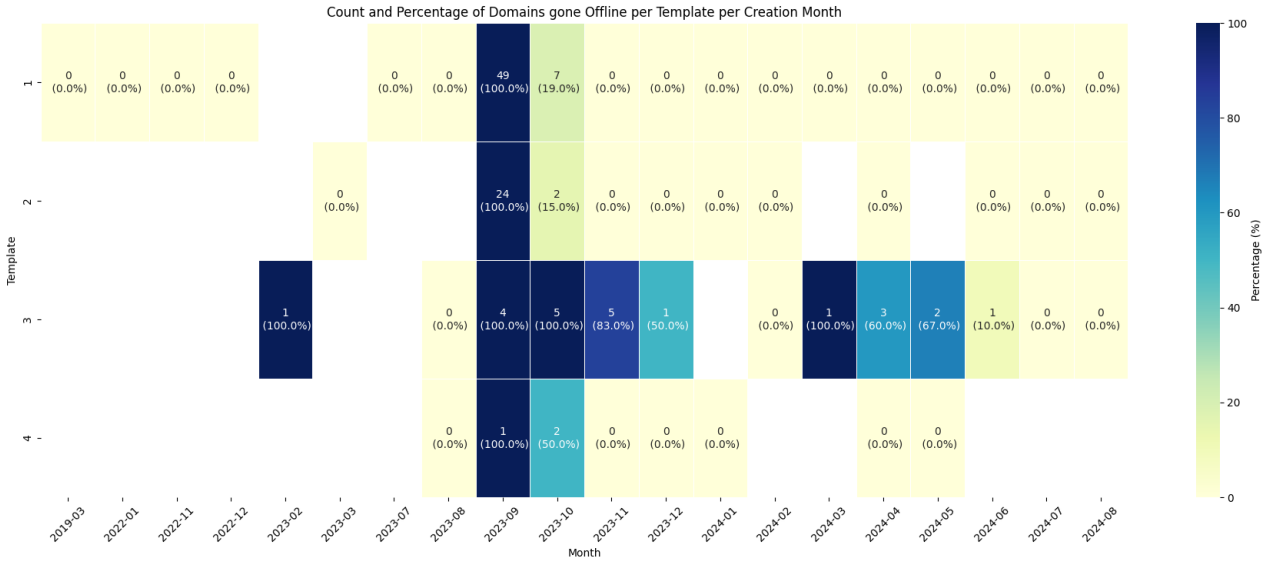


Figure 7.2: Count and percentage of domains gone offline per template per month.

Using the collected data, the lifespan of a website is calculated for all domains that have gone offline between September 2nd 2024 and October 11th 2024. This lifespan is defined as the last day the website was online minus the creation date of the domain. Figure 7.3 shows the distribution of the lifespan lengths. A high peak is witnessed at the 365 day mark: exactly one year. This indicates websites are going offline due to a contract that expires. In this case, observations that are done during the measurements confirm that this was often the expiration of the domain registry.

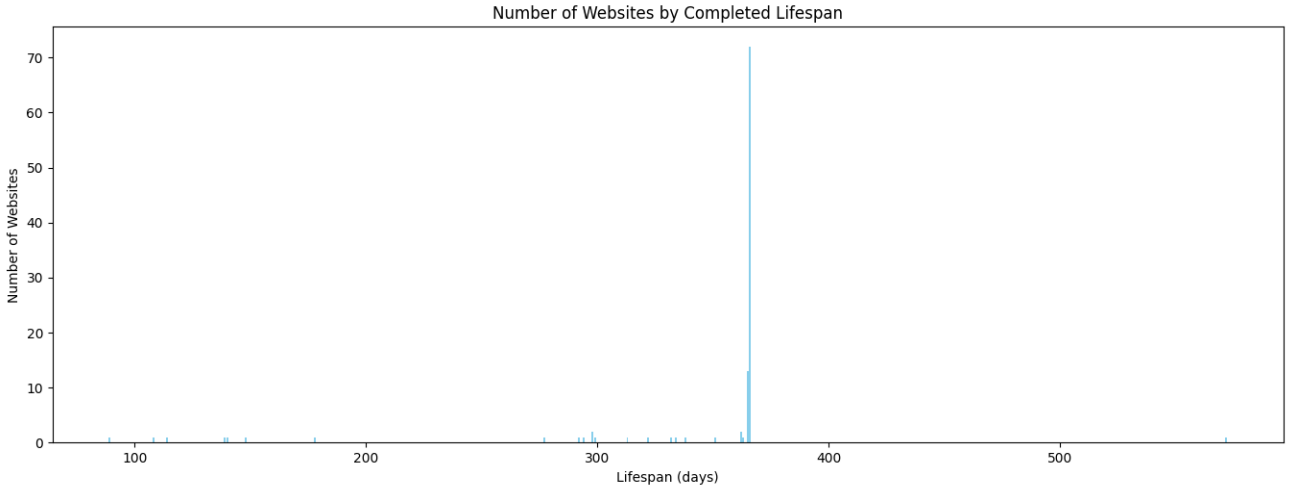


Figure 7.3: Number of websites by total lifespan.

It can thus be concluded that of all websites that have gone offline, they are most likely to go offline after exactly one year. However, in order to generalize this statement, the lifespan of all websites that are still online also need to be analyzed. This is done by computing the current days online, which is defined as October 11th 2024 minus the creation date of the domain. If the websites go offline after a year, then we expect the majority of the websites that are still online to be younger than one year. The plot is shown in figure 7.4.

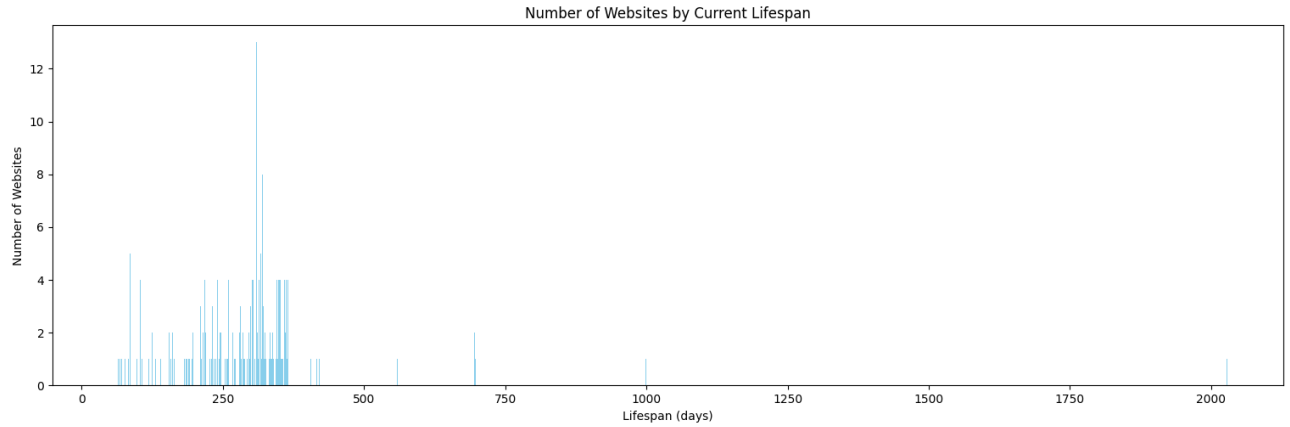
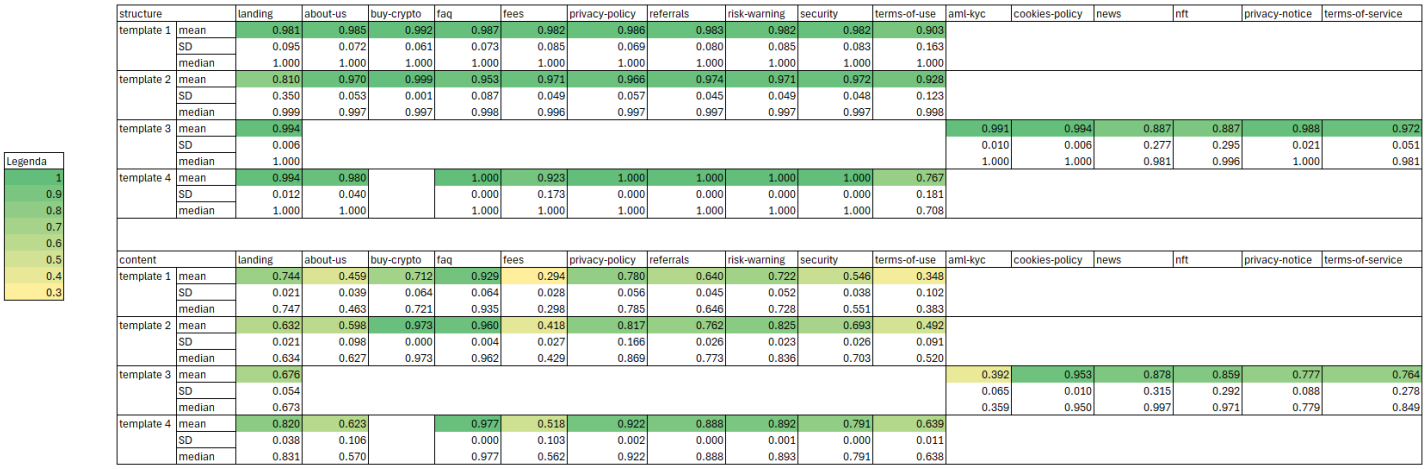


Figure 7.4: Number of websites by current lifespan.

From this plot it can be seen that the completed lifespan graph is indeed representative. There is high activity in the region up to 365 days, which means that most websites that are still online are younger than 365 days and are likely to also go offline at 365 days. Only a few domains are currently older than 365 days. However, this does not necessarily mean that the templates have the same age as these domains. A domain could have been repurposed: the original content was not the same as the current content. This could have happened at any point between the domain creation date and the labeling step of this research. This means that no conclusions can be drawn from the distribution about the age of the templates.



structure		landing	about-us	buy-crypto	faq	fees	privacy-policy	referrals	risk-warning	security	terms-of-use	aml-kyc	cookies-policy	news	nft	privacy-notice	terms-of-service
template 1	mean	0.981	0.985	0.992	0.987	0.982	0.986	0.983	0.982	0.982	0.903						
	SD	0.095	0.072	0.061	0.073	0.085	0.069	0.080	0.085	0.083	0.163						
	median	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000						
template 2	mean	0.810	0.970	0.999	0.953	0.971	0.966	0.974	0.971	0.972	0.928						
	SD	0.350	0.053	0.001	0.087	0.049	0.057	0.045	0.049	0.048	0.123						
	median	0.999	0.997	0.997	0.998	0.996	0.997	0.997	0.997	0.997	0.998						
template 3	mean	0.994										0.991	0.994	0.887	0.887	0.988	0.972
	SD	0.006										0.010	0.006	0.277	0.295	0.021	0.051
	median	1.000										1.000	1.000	0.981	0.996	1.000	0.981
template 4	mean	0.994	0.980		1.000	0.923	1.000	1.000	1.000	1.000	0.767						
	SD	0.012	0.040		0.000	0.173	0.000	0.000	0.000	0.000	0.181						
	median	1.000	1.000		1.000	1.000	1.000	1.000	1.000	1.000	0.708						

Legenda
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3

content		landing	about-us	buy-crypto	faq	fees	privacy-policy	referrals	risk-warning	security	terms-of-use	aml-kyc	cookies-policy	news	nft	privacy-notice	terms-of-service
template 1	mean	0.744	0.459	0.712	0.929	0.294	0.780	0.640	0.722	0.546	0.348						
	SD	0.021	0.039	0.064	0.064	0.028	0.056	0.045	0.052	0.038	0.102						
	median	0.747	0.463	0.721	0.935	0.298	0.785	0.646	0.728	0.551	0.383						
template 2	mean	0.632	0.598	0.973	0.960	0.418	0.817	0.762	0.825	0.693	0.492						
	SD	0.021	0.098	0.000	0.004	0.027	0.166	0.026	0.023	0.026	0.091						
	median	0.634	0.627	0.973	0.962	0.429	0.869	0.773	0.836	0.703	0.520						
template 3	mean	0.676										0.392	0.953	0.878	0.859	0.777	0.764
	SD	0.054										0.065	0.010	0.315	0.292	0.088	0.278
	median	0.673										0.359	0.950	0.997	0.971	0.779	0.849
template 4	mean	0.820	0.623		0.977	0.518	0.922	0.888	0.892	0.791	0.639						
	SD	0.038	0.106		0.000	0.103	0.002	0.000	0.001	0.000	0.011						
	median	0.831	0.570		0.977	0.562	0.922	0.888	0.893	0.791	0.638						

Figure 7.5: Mean, standard deviation, and median per template page.

7.2. Intra-template similarity

In figure 7.5, the means, standard deviations and medians of the 72 constructed intra-template similarity tables are shown. The means are colored from yellow to green. Important to note is that the *buy-crypto* page similarities of templates 1 and 2 are based on less webpages than the total number of websites within the template. This is due to this webpage being present on some websites within the templates, but not existing on others. Additionally, template 3 does not share the page names with the other three templates. This suggests that this template is different from templates 1,2 and 4. However, this will be reviewed in more detail in the inter-template analysis.

In general, it can be said that low mean similarity values indicate more diversity within a template and high mean similarities indicate homogeneity within a template. However, a change in a short text can have a larger impact on the similarity score than an equal change in a longer text. Furthermore, a relatively large difference between the median and the mean indicates that the data is skewed in a single direction. This can be caused by one or more outliers that have influenced the mean degree of similarity. An example of this is the structural similarity of the *terms of use* page of template 1. The mean value is 0.10 lower than the median value of 1. When analyzing the table which these statistics are based on, it becomes clear that about 15% of the values is below the mean.

The structural similarity is overall very high. Thus, it can be concluded that websites use the same HTML structure as the rest of the websites within the template and hardly deviate from it. The *terms of use* page of template 4 is the only outlier with a mean value of 0.77. This indicates that this page has the most diversity with regards to HTML structure.

The textual content statistical values are more spread out and there are therefore no significant outliers. However, the *fees*, *terms of use* and *aml-kyc* (anti money laundering - know your customer) pages have the lowest mean similarity and the *FAQ* page has the highest mean similarity.

When looking at the original similarity tables that these values are based on, it was not necessarily the high or low similarities that stood out. It was the degree of deviation within the template. In table 7.5 this is captured by the standard deviation (SD) value. A high standard deviation can either mean that a small portion of the websites have made drastic changes with regards to an original template or that many websites have made some smaller changes. These smaller changes could either come from individual choices of website owners or from more coordinated deviations. Important to note is the variation in sample size between the templates. An outlier with template 1 will have less impact on the average mean than an equal outlier with template 4.

Table 7.1 shows all webpages with a textual content standard deviation of over 0.10. They are reviewed in order to establish the reason for the relatively high standard deviation. It can be seen for template 4 *fees* that a small amount of moderate outliers already causes a standard deviation of over 0.10. This can be explained by the low number of websites with template 4. The high standard deviation for webpages *news*, *nft* and *terms of service* of template 3 are all due to having two pages with vastly different content. These outliers are more extreme than the outliers for the template 4 *fees* webpages.

Table 7.1: Standard deviation review of textual content similarities

Template/page	SD	Reason
1/terms of use	**0.102**	moderate amount of moderate outliers
2/privacy policy	**0.166**	small amount of extreme outliers
3/news	**0.315**	small amount of extreme outliers
3/nft	**0.292**	small amount of extreme outliers (different from 3/news outliers)
3/terms of service	**0.278**	small amount of extreme outliers (1 outlier shared with 3/nft)
4/fees	**0.103**	small amount of moderate outliers

The statistics based on the full website similarities are shown in figure 7.6. It shows that template 1 has the lowest textual content similarities overall, but also has the highest structural content similarities.

Legenda
1
0.9
0.8
0.7
0.6

Average similarities		template 1	template 2	template 3	template 4
structural	mean	0.973	0.929	0.959	0.963
	SD	0.088	0.122	0.060	0.029
	median	1.000	0.997	0.985	1.000
content	mean	0.615	0.674	0.757	0.786
	SD	0.042	0.042	0.094	0.018
	median	0.625	0.705	0.792	0.018

Figure 7.6: Mean and standard deviation per template average.

7.2.1. Influence of domain name

The content similarity analysis is repeated with the extra processing step of replacing each instance of the domain name with a placeholder that is equal for all websites and for the extra step of removing each instance of the domain name. Replacing the domain names causes the average content similarities to be much higher. Even higher values are reached when removing the domain names from the analysis entirely. Template 1 has the largest delta between the average similarity values when domain names are present and when they are either replaced with a placeholder or removed entirely.

When looking at the similarity tables of template 1 it can concluded that the initial hypothesis was correct. For the pages of template 1 the similarity scores were often equal to 1, which means that webpages were perfect copies apart from the domain name. This was true for both replacing and removing the domain names, but more often for removing the domain name, which indicates that the only difference between certain websites is a mismatch in domain name occurrences. Templates 2, 3 and 4 show some more diversity than template 1, but they too show an increase in perfectly similar webpages. Thus, the hypothesis that there are no differences between websites apart from the domain name is only partially true.

Legenda
1
0.9
0.8
0.7
0.6

Average content similarities		template 1	template 2	template 3	template 4
with domains	mean	0.615	0.674	0.757	0.786
	SD	0.042	0.042	0.094	0.018
	median	0.625	0.705	0.792	0.018
with domain placeholder	mean	0.907	0.827	0.853	0.905
	SD	0.111	0.100	0.120	0.059
	median	0.929	0.823	0.884	0.903
with domains removed	mean	0.946	0.869	0.872	0.953
	SD	0.095	0.097	0.123	0.042
	median	0.993	0.869	0.903	0.964

Figure 7.7: Comparison of average content similarity values with domain names, with domain placeholder, and with domain names removed.

These methods also unveiled one striking outlier in template 1. Further investigation suggested that this website was not scraped correctly and that the landing page was repeatedly saved instead of the sub pages. This is likely due to a temporary internet connection interruption. This website, with ID 1-173, was consequently removed from further analyses.

7.3. Clustering

As described in the methodology, in order to perform the inter-template analysis the templates need to be separated into a core cluster and a remainder. If a website is taken from the core of a cluster, then it is a representative website for this cluster. These steps have been precisely documented, since the thresholds were chosen based on the internal clustering behavior of a template. Deciding at which thresholds to stop the clustering is somewhat discretionary, but generally the clustering stopped whenever there were no more non-trivial splits left to make. Below, the taken steps and final core clusters can be found. The websites are represented by an index number.

7.3.1. Clustering of template 1

Table 7.2 describes the steps taken in order to get to the core cluster of template 1. There are 275 websites with template 1 in the dataset, since one website has been disregarded. The resulting core cluster can be seen in table 7.3.

Table 7.2: Cluster core approximation steps for template 1

Struc/Cont	Threshold	Eliminated items template 1 (size=44)
Structure	0.8	[0,75,82,95,257]
Structure	0.95	[58]
Content	0.6	[73]
Content	0.62	[3,6,11,22,23,26,31,42,43,46,55,67,80,85,93,98,101,109,143,167, 170,174,175,180,193,212,237,241,242,250,255,268,271,275]
Structure	0.99	[117,244,272]

Table 7.3: Cluster core of template 1

Cluster core of template 1 (size = 231)
{1, 2, 4, 5, 7, 8, 9, 10, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 24, 25, 27, 28, 29, 30, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 44, 45, 47, 48, 49, 50, 51, 52, 53, 54, 56, 57, 59, 60, 61, 62, 63, 64, 65, 66, 68, 69, 70, 71, 72, 74, 76, 77, 78, 79, 81, 83, 84, 86, 87, 88, 89, 90, 91, 92, 94, 96, 97, 99, 100, 102, 103, 104, 105, 106, 107, 108, 110, 111, 112, 113, 114, 115, 116, 118, 119, 120, 121, 122, 123, 124, 125, 126, 127, 128, 129, 130, 131, 132, 133, 134, 135, 136, 137, 138, 139, 140, 141, 142, 144, 145, 146, 147, 148, 149, 150, 151, 152, 153, 154, 155, 156, 157, 158, 159, 160, 161, 162, 163, 164, 165, 166, 168, 169, 171, 172, 176, 177, 178, 179, 181, 182, 183, 184, 185, 186, 187, 188, 189, 190, 191, 192, 194, 195, 196, 197, 198, 199, 200, 201, 202, 203, 204, 205, 206, 207, 208, 209, 210, 211, 213, 214, 215, 216, 217, 218, 219, 220, 221, 222, 223, 224, 225, 226, 227, 228, 229, 230, 231, 232, 233, 234, 235, 236, 238, 239, 240, 243, 245, 246, 247, 248, 249, 251, 252, 253, 254, 256, 258, 259, 260, 261, 262, 263, 264, 265, 266, 267, 269, 270, 273, 274}

7.3.2. Clustering of template 2

The same steps are performed for the 66 website of template 2. These results can be seen in table 7.4. When finding the core for cluster 2, one large cluster and one smaller cluster are discovered. Both clusters are structurally identical to each other. The items have textual content similarity scores that are high within their cluster, but lower across the two clusters. Since this is an interesting partitioning, both cores are included in the inter-template analysis. The large cluster is from now on referred to as template 2a and the smaller cluster as template 2b.

Table 7.4: Cluster core approximation steps for template 2

Struc/Cont	Threshold	Eliminated items template 2 (size=19)
Structure	0.8	[15,22,24,30,35,49,55,56]
Structure	0.99	[1,2,12,18,45]
Content	0.7	[20,25,27,58,60]
Structure	0.9999999999999999	[3]

Table 7.5: Cluster cores of template 2

	Cluster cores (sizes: 2a=35 2b=12)
Template 2a	{6, 7, 8, 9, 10, 11, 13, 14, 16, 19, 21, 23, 26, 31, 32, 33, 36, 37, 38, 39, 40, 41, 43, 44, 46, 47, 48, 50, 52, 53, 54, 59, 62, 63, 64}
Template 2b	{0, 4, 5, 17, 28, 29, 34, 42, 51, 57, 61, 65}

7.3.3. Clustering of template 3

Template 3 does not show such clear outliers when using the division-based clustering method for the structural similarity. The 31 websites can be clustered in multiple ways in which the minimum mutual similarity is 0.8. Therefore the intersection of these clusters is considered as core instead. The items that did not belong to the intersection are eliminated. This approach is verified by the computation of the content similarity clusters with threshold 0.74, which shows items {0, 9, 12, 14, 18, 23} as anomalies. The core cluster is shown in 7.6.

Table 7.6: Cluster core of template 3

Cluster core of template 3 (size=25)
{1, 2, 3, 4, 5, 6, 7, 8, 10, 11, 13, 15, 16, 17, 19, 20, 21, 22, 24, 25, 26, 27, 28, 29, 30}

7.3.4. Clustering of template 4

For the 23 websites of template 4, the intersection approach works most effectively. The intersections are taken of the sets which result in a structure clustering with threshold 0.9. The items that do not belong to the intersection are eliminated. These are {2,7,10}. No further meaningful splits based on the textual content similarities could be made. This results in the core cluster seen in 7.7.

Table 7.7: Cluster core of template 4

Cluster core of template 4 (size=20)
{0, 1, 3, 4, 5, 6, 8, 9, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22}

7.4. Inter-template similarity

From the cluster cores, 5 websites are randomly selected for inter-template analysis, one per template for templates 1, 2a, 2b, 3, and 4. The selected websites are shown in table 7.8.

Table 7.8: Websites chosen for the inter-template analysis

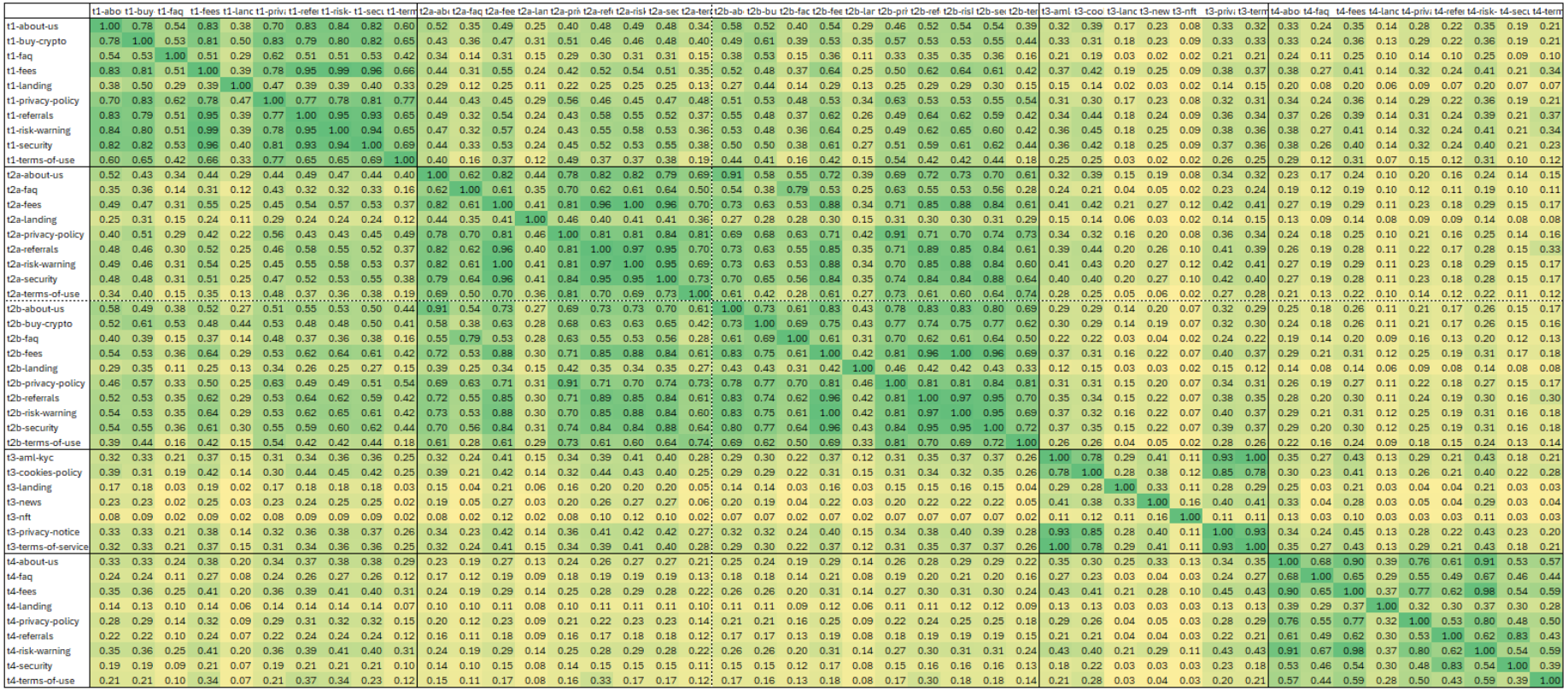
Template	ID	Website Name
1	71	*cryptofamas.com*
2a	8	*bitmantos.com*
2b	51	*salorex.com*
3	30	*usetradex.com*
4	13	*ganbitc.com*

All webpages from the 5 representative websites are compared pairwise, both in terms of structure and content. The pairwise structural similarities are shown in figure 7.8. The colors of the table depict their similarity score, for which green means high similarity and yellow means a lower similarity. It can be seen from the green squares of cells around the diagonal that the structure of the website is quite consistent within a template, although template 3 is an exception to this observation.

Overall, the landing pages have the most unique HTML structure. A logical explanation is that the landing page is seen first and tries to leave a good impression.

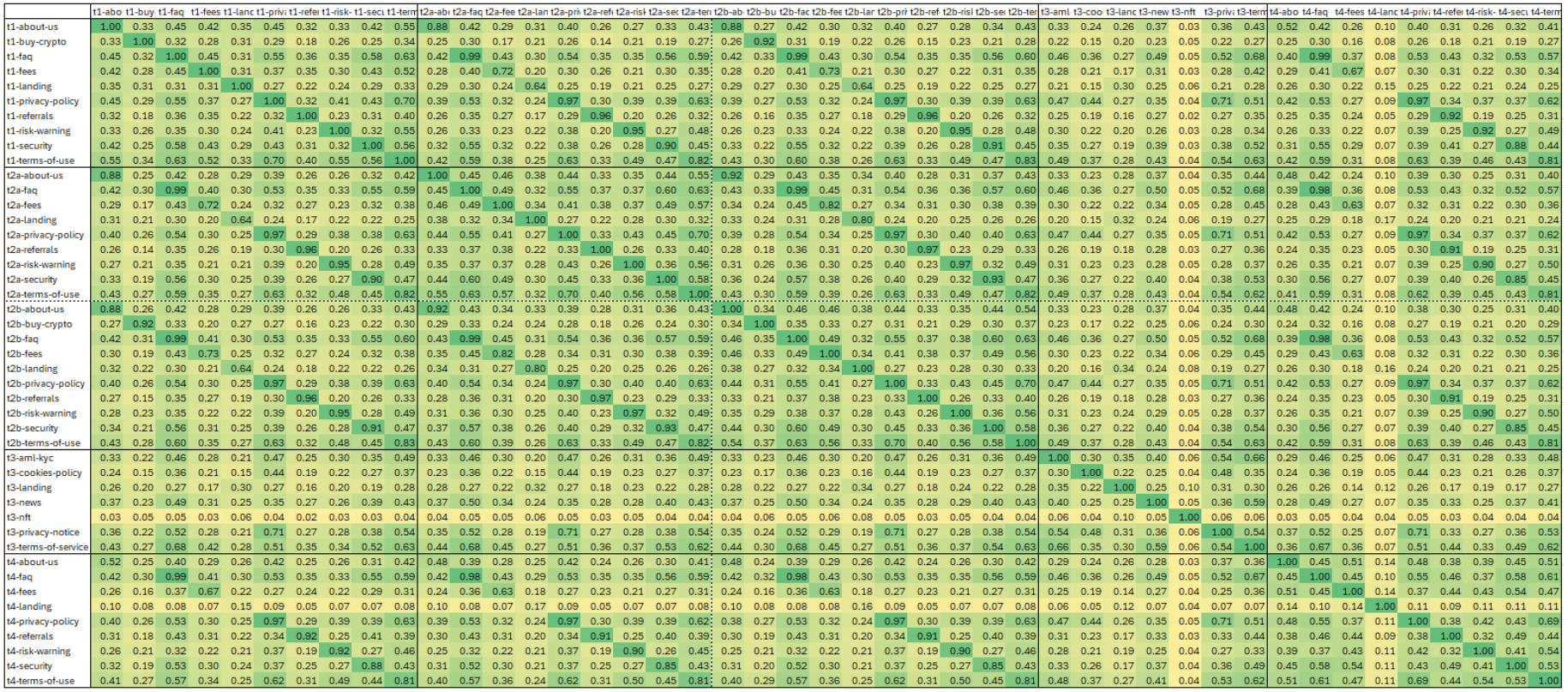
Figure 7.9 shows the pairwise textual content similarities of the five representative websites. Again, green represents a high textual content similarity and yellow represents a lower similarity. Since all webpages are paired against all other webpages, it is not expected that everything will yield high similarities. However, it is expected to yield a high similarity when comparing webpages with a similar topic from different templates. The green diagonal lines of cells indicate that indeed the content is shared among templates. The lack of diagonal green lines in the columns and rows of template 3 indicate that template 3 does not share as much content with the other 3 templates. Templates 1 and 2 match the most with each other, whereas template 4 seems to only share textual content with templates 1 and 2 on some of the webpages.

The two websites from template 2 still share more content between themselves than with the other templates. This can be concluded since the similarity values between their webpage pairs are higher than combinations with other templates.



	t1-abo	t1-buy	t1-faq	t1-fees	t1-lanc	t1-priv	t1-refe	t1-risk-	t1-sect	t1-term	t2a-ab	t2a-faq	t2a-fee	t2a-lan	t2a-pri	t2a-refe	t2a-risk	t2a-sec	t2a-ter	t2b-ab	t2b-bu	t2b-fac	t2b-fee	t2b-lar	t2b-pri	t2b-ref	t2b-risk	t2b-ser	t2b-ter	t3-aml	t3-coo	t3-lanc	t3-new	t3-nft	t3-priv	t3-term	t4-abo	t4-faq	t4-fees	t4-lanc	t4-priv	t4-refe	t4-risk-	t4-sect	t4-term
t1-about-us	1.00	0.78	0.54	0.83	0.38	0.70	0.83	0.84	0.82	0.60	0.52	0.35	0.49	0.25	0.40	0.48	0.49	0.48	0.34	0.58	0.52	0.40	0.54	0.29	0.46	0.52	0.54	0.54	0.39	0.32	0.39	0.17	0.23	0.08	0.33	0.32	0.33	0.24	0.35	0.14	0.28	0.22	0.35	0.19	0.21
t1-buy-crypto	0.78	1.00	0.53	0.81	0.50	0.83	0.79	0.80	0.82	0.65	0.43	0.36	0.47	0.31	0.51	0.46	0.46	0.48	0.40	0.49	0.61	0.39	0.53	0.35	0.57	0.53	0.53	0.55	0.44	0.33	0.31	0.18	0.23	0.09	0.33	0.33	0.33	0.24	0.36	0.13	0.29	0.22	0.36	0.19	0.21
t1-faq	0.54	0.53	1.00	0.51	0.29	0.62	0.51	0.51	0.53	0.42	0.34	0.14	0.31	0.15	0.29	0.30	0.31	0.31	0.15	0.38	0.53	0.15	0.36	0.11	0.33	0.35	0.35	0.36	0.16	0.21	0.19	0.03	0.02	0.02	0.21	0.21	0.24	0.11	0.25	0.10	0.14	0.10	0.25	0.09	0.10
t1-fees	0.83	0.81	0.51	1.00	0.39	0.78	0.95	0.99	0.96	0.66	0.44	0.31	0.55	0.24	0.42	0.52	0.54	0.51	0.35	0.52	0.48	0.37	0.64	0.25	0.50	0.62	0.64	0.61	0.42	0.37	0.42	0.19	0.25	0.09	0.38	0.37	0.38	0.27	0.41	0.14	0.32	0.24	0.41	0.21	0.34
t1-landing	0.38	0.50	0.29	0.39	1.00	0.47	0.39	0.39	0.40	0.33	0.29	0.12	0.25	0.11	0.22	0.25	0.25	0.25	0.13	0.27	0.44	0.14	0.29	0.13	0.25	0.29	0.29	0.30	0.15	0.15	0.14	0.02	0.03	0.02	0.14	0.15	0.20	0.08	0.20	0.06	0.09	0.07	0.20	0.07	0.07
t1-privacy-policy	0.70	0.83	0.62	0.78	0.47	1.00	0.77	0.78	0.81	0.77	0.44	0.43	0.45	0.29	0.56	0.46	0.45	0.47	0.48	0.51	0.53	0.48	0.53	0.34	0.63	0.53	0.53	0.55	0.54	0.31	0.30	0.17	0.23	0.08	0.32	0.31	0.34	0.24	0.36	0.14	0.29	0.22	0.36	0.19	0.21
t1-referrals	0.83	0.79	0.51	0.95	0.39	0.77	1.00	0.95	0.93	0.65	0.49	0.32	0.54	0.24	0.43	0.58	0.55	0.52	0.37	0.55	0.48	0.37	0.62	0.26	0.49	0.64	0.62	0.59	0.42	0.34	0.44	0.18	0.24	0.09	0.36	0.34	0.37	0.26	0.39	0.14	0.31	0.24	0.39	0.21	0.37
t1-risk-warning	0.84	0.80	0.51	0.99	0.39	0.78	0.95	1.00	0.94	0.65	0.47	0.32	0.57	0.24	0.43	0.55	0.58	0.53	0.36	0.53	0.48	0.36	0.64	0.25	0.49	0.62	0.65	0.60	0.42	0.36	0.45	0.18	0.25	0.09	0.38	0.36	0.38	0.27	0.41	0.14	0.32	0.24	0.41	0.21	0.34
t1-security	0.82	0.82	0.53	0.96	0.40	0.81	0.93	0.94	1.00	0.69	0.44	0.33	0.53	0.24	0.45	0.52	0.53	0.55	0.38	0.50	0.50	0.38	0.61	0.27	0.51	0.59	0.61	0.62	0.44	0.36	0.42	0.18	0.25	0.09	0.37	0.36	0.38	0.26	0.40	0.14	0.32	0.24	0.40	0.21	0.23
t1-terms-of-use	0.60	0.65	0.42	0.66	0.33	0.77	0.65	0.65	0.69	1.00	0.40	0.16	0.37	0.12	0.49	0.37	0.37	0.38	0.19	0.44	0.41	0.16	0.42	0.15	0.54	0.42	0.42	0.44	0.18	0.25	0.25	0.03	0.02	0.02	0.26	0.25	0.29	0.12	0.31	0.07	0.15	0.12	0.31	0.10	0.12
t2a-about-us	0.52	0.43	0.34	0.44	0.29	0.44	0.49	0.47	0.44	0.40	1.00	0.62	0.82	0.44	0.78	0.82	0.82	0.79	0.69	0.91	0.58	0.55	0.72	0.39	0.69	0.72	0.73	0.70	0.61	0.32	0.39	0.15	0.19	0.08	0.34	0.32	0.23	0.17	0.24	0.10	0.20	0.16	0.24	0.14	0.15
t2a-faq	0.35	0.36	0.14	0.31	0.12	0.43	0.32	0.32	0.33	0.16	0.62	1.00	0.61	0.35	0.70	0.62	0.61	0.64	0.50	0.54	0.38	0.79	0.53	0.25	0.63	0.55	0.53	0.56	0.28	0.24	0.21	0.04	0.05	0.02	0.23	0.24	0.19	0.12	0.19	0.10	0.12	0.11	0.19	0.10	0.11
t2a-fees	0.49	0.47	0.31	0.55	0.25	0.45	0.54	0.57	0.53	0.37	0.82	0.61	1.00	0.41	0.81	0.96	1.00	0.96	0.70	0.73	0.63	0.53	0.88	0.34	0.71	0.85	0.88	0.84	0.61	0.41	0.42	0.21	0.27	0.12	0.42	0.41	0.27	0.19	0.29	0.11	0.23	0.18	0.29	0.15	0.17
t2a-landing	0.25	0.31	0.15	0.24	0.11	0.29	0.24	0.24	0.24	0.12	0.44	0.35	0.41	1.00	0.46	0.40	0.41	0.41	0.36	0.27	0.28	0.28	0.30	0.15	0.31	0.30	0.30	0.31	0.29	0.15	0.14	0.06	0.03	0.02	0.14	0.15	0.13	0.09	0.14	0.08	0.09	0.09	0.14	0.08	0.08
t2a-privacy-policy	0.40	0.51	0.29	0.42	0.22	0.56	0.43	0.43	0.45	0.49	0.78	0.70	0.81	0.46	1.00	0.81	0.81	0.84	0.81	0.69	0.68	0.63	0.71	0.42	0.91	0.71	0.70	0.74	0.73	0.34	0.32	0.16	0.20	0.08	0.36	0.34	0.24	0.18	0.25	0.10	0.21	0.16	0.25	0.14	0.16
t2a-referrals	0.48	0.46	0.30	0.52	0.25	0.46	0.58	0.55	0.52	0.37	0.82	0.62	0.96	0.40	0.81	1.00	0.97	0.95	0.70	0.73	0.63	0.55	0.85	0.35	0.71	0.89	0.85	0.84	0.61	0.39	0.44	0.20	0.26	0.10	0.41	0.39	0.26	0.19	0.28	0.11	0.22	0.17	0.28	0.15	0.33
t2a-risk-warning	0.49	0.46	0.31	0.54	0.25	0.45	0.55	0.58	0.53	0.37	0.82	0.61	1.00	0.41	0.81	0.97	1.00	0.95	0.69	0.73	0.63	0.53	0.88	0.34	0.70	0.85	0.88	0.84	0.60	0.41	0.43	0.20	0.27	0.12	0.42	0.41	0.27	0.19	0.29	0.11	0.23	0.18	0.29	0.15	0.17
t2a-security	0.48	0.48	0.31	0.51	0.25	0.47	0.52	0.53	0.55	0.38	0.79	0.64	0.96	0.41	0.84	0.95	0.95	1.00	0.73	0.70	0.65	0.56	0.84	0.35	0.74	0.84	0.84	0.88	0.64	0.40	0.40	0.20	0.27	0.10	0.42	0.40	0.27	0.19	0.28	0.11	0.23	0.18	0.28	0.15	0.17
t2a-terms-of-use	0.34	0.40	0.15	0.35	0.13	0.48	0.37	0.36	0.38	0.19	0.69	0.50	0.70	0.36	0.81	0.70	0.69	0.73	1.00	0.61	0.42	0.28	0.61	0.27	0.73	0.61	0.60	0.64	0.74	0.28	0.25	0.05	0.06	0.02	0.27	0.28	0.21	0.13	0.22	0.10	0.14	0.12	0.22	0.11	0.12
t2b-about-us	0.58	0.49	0.38	0.52	0.27	0.51	0.55	0.53	0.50	0.44	0.91	0.54	0.73	0.27	0.69	0.73	0.73	0.70	0.61	1.00	0.73	0.61	0.83	0.43	0.78	0.83	0.83	0.80	0.69	0.29	0.29	0.14	0.20	0.07	0.32	0.29	0.25	0.18	0.26	0.11	0.21	0.17	0.26	0.15	0.17
t2b-buy-crypto	0.52	0.61	0.53	0.48	0.44	0.53	0.48	0.48	0.50	0.41	0.58	0.38	0.63	0.28	0.68	0.63	0.63	0.65	0.42	0.73	1.00	0.69	0.75	0.43	0.77	0.74	0.75	0.77	0.62	0.30	0.29	0.14	0.19	0.07	0.32	0.30	0.24	0.18	0.26	0.11	0.21	0.17	0.26	0.15	0.16
t2b-faq	0.40	0.39	0.15	0.37	0.14	0.48	0.37	0.36	0.38	0.16	0.55	0.79	0.53	0.28	0.63	0.55	0.53	0.56	0.28	0.61	0.69	1.00	0.61	0.31	0.70	0.62	0.61	0.64	0.50	0.22	0.22	0.03	0.04	0.02	0.24	0.22	0.19	0.14	0.20	0.09	0.16	0.13	0.20	0.12	0.13
t2b-fees	0.54	0.53	0.36	0.64	0.29	0.53	0.62	0.64	0.61	0.42	0.72	0.53	0.88	0.30	0.71	0.85	0.88	0.84	0.61	0.83	0.75	0.61	1.00	0.42	0.81	0.96	1.00	0.96	0.69	0.37	0.31	0.16	0.22	0.07	0.40	0.37	0.29	0.21	0.31	0.12	0.25	0.19	0.31	0.17	0.18
t2b-landing	0.29	0.35	0.11	0.25	0.13	0.34	0.26	0.25	0.27	0.15	0.39	0.25	0.34	0.15	0.42	0.35	0.34	0.35	0.27	0.43	0.43	0.31	0.42	1.00	0.46	0.42	0.42	0.43	0.33	0.12	0.15	0.03	0.03	0.02	0.15	0.12	0.14	0.08	0.14	0.06	0.09	0.08	0.14	0.08	0.08
t2b-privacy-policy	0.46	0.57	0.33	0.50	0.25	0.63	0.49	0.49	0.51	0.54	0.69	0.63	0.71	0.31	0.91	0.71	0.70	0.74	0.73	0.78	0.77	0.70	0.81	0.46	1.00	0.81	0.81	0.84	0.81	0.31	0.31	0.15	0.20	0.07	0.34	0.31	0.26	0.19	0.27	0.11	0.22	0.18	0.27	0.15	0.17
t2b-referrals	0.52	0.53	0.35	0.62	0.29	0.53	0.64	0.62	0.59	0.42	0.72	0.55	0.85	0.30	0.71	0.89	0.85	0.84	0.61	0.83	0.74	0.62	0.96	0.42	0.81	1.00	0.97	0.95	0.70	0.35	0.34	0.15	0.22	0.07	0.38	0.35	0.28	0.20	0.30	0.11	0.24	0.19	0.30	0.16	0.30
t2b-risk-warning	0.54	0.53	0.35	0.64	0.29	0.53	0.62	0.65	0.61	0.42	0.73	0.53	0.88	0.30	0.70	0.85	0.88	0.84	0.60	0.83	0.75	0.61	1.00	0.42	0.81	0.97	1.00	0.95	0.69	0.37	0.32	0.16	0.22	0.07	0.40	0.37	0.29	0.21	0.31	0.12	0.25	0.19	0.31	0.16	0.18
t2b-security	0.54	0.55	0.36	0.61	0.30	0.55	0.59	0.60	0.62	0.44	0.70	0.56	0.84	0.31	0.74	0.84	0.84	0.88	0.64	0.80	0.77	0.64	0.96	0.43	0.84	0.95	0.95	1.00	0.72	0.37	0.35	0.15	0.22	0.07	0.39	0.37	0.29	0.20	0.30	0.12	0.25	0.19	0.31	0.16	0.18
t2b-terms-of-use	0.39	0.44	0.16	0.42	0.15	0.54	0.42	0.42	0.44	0.18	0.61	0.28	0.61	0.29	0.73	0.61	0.60	0.64	0.74	0.69	0.62	0.50	0.69	0.33	0.81	0.70	0.69	0.72	1.00	0.26	0.26	0.04	0.05	0.02	0.28	0.26	0.22	0.16	0.24	0.09	0.18	0.15	0.24	0.13	0.14
t3-aml-kyc	0.32	0.33	0.21	0.37	0.15	0.31	0.34	0.36	0.36	0.25	0.32	0.24	0.41	0.15	0.34	0.39	0.41	0.40	0.28	0.29	0.30	0.22	0.37	0.12	0.31	0.35	0.37	0.37	0.26	1.00	0.78	0.29	0.41	0.11	0.93	1.00	0.35	0.27	0.43	0.13	0.29	0.21	0.43	0.18	0.21
t3-cookies-policy	0.39	0.31	0.19	0.42	0.14	0.30	0.44	0.45	0.42	0.25	0.39	0.21	0.42	0.14	0.32	0.44	0.43	0.40	0.25	0.29	0.29	0.22	0.31	0.15	0.31	0.34	0.32	0.35	0.26	0.78	1.00	0.28	0.38	0.12	0.85	0.78	0.30	0.23	0.41	0.13	0.26	0.21	0.40	0.22	0.28
t3-landing	0.17	0.18	0.03	0.19	0.02	0.17	0.18	0.18	0.18	0.03	0.15	0.04	0.21	0.06	0.16	0.20	0.20	0.20	0.05	0.14	0.14	0.03	0.16	0.03	0.15	0.15	0.16	0.15	0.04	0.29	0.28	1.00	0.33	0.11	0.28	0.29	0.25	0.03	0.21	0.03	0.04	0.04	0.21	0.03	0.03
t3-news	0.23	0.23	0.02	0.25	0.03	0.23	0.24	0.25	0.25	0.02	0.19	0.05	0.27	0.03	0.20	0.26	0.27	0.27	0.06	0.20	0.19	0.04	0.22	0.03	0.20	0.22	0.22	0.22	0.05	0.41	0.38	0.33	1.00	0.16	0.40	0.41	0.33	0.04	0.28	0.03	0.05	0.04	0.29	0.03	0.04
t3-nft	0.08	0.09	0.02	0.09	0.02	0.08	0.09	0.09	0.09	0.02	0.08	0.02	0.12	0.02	0.08	0.10	0.12	0.10	0.02	0.07	0.07	0.02	0.07	0.02	0.07	0.07	0.07	0.07	0.02	0.11	0.12	0.11	0.16	1.00	0.11	0.11	0.13	0.03	0.10	0.03	0.03	0.03	0.11	0.03	0.03
t3-privacy-notice	0.33	0.33	0.21	0.38	0.14	0.32	0.36	0.38	0.37	0.26	0.34	0.23	0.42	0.14	0.36	0.41	0.42	0.42	0.27	0.32	0.32	0.24	0.40	0.15	0.34	0.38	0.40	0.39	0.28	0.93	0.85	0.28	0.40	0.11	1.00	0.93	0.34	0.24	0.45	0.13	0.28	0.22	0.43	0.23	0.20
t3-terms-of-service	0.32	0.33	0.21	0.37	0.15	0.31	0.34	0.36	0.36	0.25	0.32	0.24	0.41	0.15	0.34	0.39	0.41	0.40	0.28	0.29	0.30	0.22	0.37	0.12	0.31	0.35	0.37	0.37	0.26	1.00	0.78	0.29	0.41	0.11	0.93	1.00	0.35	0.27	0.43	0.13	0.29	0.21	0.43	0.18	0.21
t4-about-us	0.33	0.33	0.24	0.38	0.20	0.34	0.37	0.38	0.38	0.29	0.23	0.19	0.27	0.13	0.24	0.26	0.27	0.27	0.21	0.25	0.24	0.19	0.29	0.14	0.26	0.28	0.29	0.29	0.22	0.35	0.30	0.25	0.33	0.13	0.34	0.35	1.00	0.68	0.90	0.39	0.76	0.61	0.91	0.53	0.57
t4-faq	0.24	0.24	0.11	0.27	0.08	0.24	0.26	0.27	0.26	0.12	0.17	0.12	0.19	0.09	0.18	0.19	0.19	0.19	0.13	0.18	0.18	0.14	0.21	0.08	0.19	0.20	0.21	0.20	0.16	0.27	0.23	0.03	0.04	0.03	0.24	0.27	0.68	1.00	0.65	0.29	0.55	0.49	0.67	0.46	0.44
t4-fees	0.35	0.36	0.25	0.41	0.20	0.36	0.39	0.41	0.40	0.31	0.24	0.19	0.29	0.14	0.25	0.28	0.29	0.28	0.22	0.26	0.26	0.20	0.31	0.14	0.27	0.30	0.31	0.30	0.24	0.43	0.41	0.21	0.28	0.10	0.45	0.43	0.90	0.65	1.00	0.37	0.77	0.62	0.98	0.54	0.59
t4-landing	0.14	0.13	0.10	0.14	0.06	0.14	0.14	0.14	0.14	0.07	0.10	0.10	0.11	0.08	0.10	0.11	0.11	0.11	0.10	0.11	0.11	0.09	0.12	0.06	0.11	0.11	0.12	0.12	0.09	0.13	0.13	0.03	0.03	0.03	0.13	0.13	0.39	0.29	0.37	1.00	0.32	0.30	0.37	0.30	0.28
t4-privacy-policy	0.28	0.29	0.14	0.32	0.09	0.29	0.31	0.32	0.32	0.15	0.20	0.12	0.23	0.09	0.21	0.22	0.23	0.23	0.14	0.21	0.21	0.16	0.25	0.09	0.22	0.24	0.25	0.25	0.18	0.29	0.26	0.04	0.05	0.03	0.28	0.29	0.76	0.55	0.77	0.32	1.00	0.53	0.80	0.48	0.50
t4-referrals	0.22	0.22	0.10	0.24	0.07	0.22	0.24	0.24	0.24	0.12	0.16	0.11	0.18	0.09	0.16	0.17	0.18	0.18	0.12	0.17	0.17	0.13	0.19	0.08	0.18	0.19	0.19	0.19	0.15	0.21	0.21	0.04	0.04	0.03	0.22	0.21	0.61	0.49	0.62	0.30	0.53	1.00	0.62	0.83	0.43
t4-risk-warning	0.35	0.36	0.25	0.41	0.20	0.36	0.39	0.41	0.40	0.31	0.24	0.19	0.29	0.14	0.25	0.28	0.29	0.28	0.22	0.26	0.26	0.20	0.31	0.14	0.27	0.30	0.31	0.31	0.24	0.43	0.40	0.21	0.29	0.11	0.43	0.43	0.91	0.67	0.98	0.37	0.80	0.62	1.00	0.54	0.59
t4-security	0.19	0.19	0.09	0.21	0.07	0.19	0.21	0.21	0.21	0.10	0.14	0.10	0.15	0.08	0.14	0.15	0.15	0.15	0.11	0.15	0.15	0.12	0.17	0.08	0.15	0.16	0.16	0.16	0.13	0.18	0.22	0.03	0.03	0.03	0.23	0.18	0.53	0.46	0.54	0.30	0.48	0.83	0.54	1.00	0.39
t4-terms-of-use	0.21	0.21	0.10	0.34	0.07	0.21	0.37	0.34	0.23	0.12	0.15	0.11	0.17	0.08	0.16	0.33	0.17	0.17	0.12	0.17	0.16	0.13	0.18	0.08	0.17	0.30	0.18	0.18	0.14	0.21	0.28	0.03	0.04	0.03	0.20	0.21	0.57	0.44	0.59	0.28	0.50	0.43	0.59	0.39	1.00

Figure 7.8: Pairwise structure similarities for representative websites.



	t1-abo	t1-buy	t1-faq	t1-fees	t1-lanc	t1-priv	t1-refe	t1-risk-	t1-sec	t1-term	t2a-ab	t2a-faq	t2a-fee	t2a-lan	t2a-pri	t2a-refe	t2a-risk	t2a-sec	t2a-ter	t2b-ab	t2b-bu	t2b-fac	t2b-fee	t2b-lar	t2b-pri	t2b-ref	t2b-risk	t2b-sec	t2b-ter	t3-aml	t3-coo	t3-lanc	t3-new	t3-nft	t3-priv	t3-term	t4-abo	t4-faq	t4-fees	t4-lanc	t4-priv	t4-refe	t4-risk-	t4-sec	t4-term
t1-about-us	1.00	0.33	0.45	0.42	0.35	0.45	0.32	0.33	0.42	0.55	0.88	0.42	0.29	0.31	0.40	0.26	0.27	0.33	0.43	0.88	0.27	0.42	0.30	0.32	0.40	0.27	0.28	0.34	0.43	0.33	0.24	0.26	0.37	0.03	0.36	0.43	0.52	0.42	0.26	0.10	0.40	0.31	0.26	0.32	0.41
t1-buy-crypto	0.33	1.00	0.32	0.28	0.31	0.29	0.18	0.26	0.25	0.34	0.25	0.30	0.17	0.21	0.26	0.14	0.21	0.19	0.27	0.26	0.92	0.31	0.19	0.22	0.26	0.15	0.23	0.21	0.28	0.22	0.15	0.20	0.23	0.05	0.22	0.27	0.25	0.30	0.16	0.08	0.26	0.18	0.21	0.19	0.27
t1-faq	0.45	0.32	1.00	0.45	0.31	0.55	0.36	0.35	0.58	0.63	0.42	0.99	0.43	0.30	0.54	0.35	0.35	0.56	0.59	0.42	0.33	0.99	0.43	0.30	0.54	0.35	0.35	0.56	0.60	0.46	0.36	0.27	0.49	0.05	0.52	0.68	0.40	0.99	0.37	0.08	0.53	0.43	0.32	0.53	0.57
t1-fees	0.42	0.28	0.45	1.00	0.31	0.37	0.35	0.30	0.43	0.52	0.28	0.40	0.72	0.20	0.30	0.26	0.21	0.30	0.35	0.28	0.20	0.41	0.73	0.21	0.30	0.27	0.22	0.31	0.35	0.28	0.21	0.17	0.31	0.03	0.28	0.42	0.29	0.41	0.67	0.07	0.30	0.31	0.22	0.30	0.34
t1-landing	0.35	0.31	0.31	0.31	1.00	0.27	0.22	0.24	0.29	0.33	0.29	0.30	0.24	0.64	0.25	0.19	0.21	0.25	0.27	0.29	0.27	0.30	0.25	0.64	0.25	0.19	0.22	0.25	0.27	0.21	0.15	0.30	0.25	0.06	0.21	0.28	0.26	0.30	0.22	0.15	0.25	0.22	0.21	0.24	0.25
t1-privacy-policy	0.45	0.29	0.55	0.37	0.27	1.00	0.32	0.41	0.43	0.70	0.39	0.53	0.32	0.24	0.97	0.30	0.39	0.39	0.63	0.39	0.27	0.53	0.32	0.24	0.97	0.30	0.39	0.39	0.63	0.47	0.44	0.27	0.35	0.04	0.71	0.51	0.42	0.53	0.27	0.09	0.97	0.34	0.37	0.37	0.62
t1-referrals	0.32	0.18	0.36	0.35	0.22	0.32	1.00	0.23	0.31	0.40	0.26	0.35	0.27	0.17	0.29	0.96	0.20	0.26	0.32	0.26	0.16	0.35	0.27	0.18	0.29	0.96	0.20	0.26	0.32	0.25	0.19	0.16	0.27	0.02	0.27	0.35	0.25	0.35	0.24	0.05	0.29	0.92	0.19	0.25	0.31
t1-risk-warning	0.33	0.26	0.35	0.30	0.24	0.41	0.23	1.00	0.32	0.55	0.26	0.33	0.23	0.22	0.38	0.20	0.95	0.27	0.48	0.26	0.23	0.33	0.24	0.22	0.38	0.20	0.95	0.28	0.48	0.30	0.22	0.20	0.26	0.03	0.28	0.34	0.26	0.33	0.22	0.07	0.39	0.25	0.92	0.27	0.49
t1-security	0.42	0.25	0.58	0.43	0.29	0.43	0.31	0.32	1.00	0.56	0.32	0.55	0.32	0.22	0.38	0.26	0.28	0.90	0.45	0.33	0.22	0.55	0.32	0.22	0.39	0.26	0.28	0.91	0.45	0.35	0.27	0.19	0.39	0.03	0.38	0.52	0.31	0.55	0.29	0.07	0.39	0.41	0.27	0.88	0.44
t1-terms-of-use	0.55	0.34	0.63	0.52	0.33	0.70	0.40	0.55	0.56	1.00	0.42	0.59	0.38	0.25	0.63	0.33	0.49	0.47	0.82	0.43	0.30	0.60	0.38	0.26	0.63	0.33	0.49	0.47	0.83	0.49	0.37	0.28	0.43	0.04	0.54	0.63	0.42	0.59	0.31	0.08	0.63	0.39	0.46	0.43	0.81
t2a-about-us	0.88	0.25	0.42	0.28	0.29	0.39	0.26	0.26	0.32	0.42	1.00	0.45	0.46	0.38	0.44	0.33	0.35	0.44	0.55	0.92	0.29	0.43	0.35	0.34	0.40	0.28	0.31	0.37	0.43	0.33	0.23	0.28	0.37	0.04	0.35	0.44	0.48	0.42	0.24	0.10	0.39	0.30	0.25	0.31	0.40
t2a-faq	0.42	0.30	0.99	0.40	0.30	0.53	0.35	0.33	0.55	0.59	0.45	1.00	0.49	0.32	0.55	0.37	0.37	0.60	0.63	0.43	0.33	0.99	0.45	0.31	0.54	0.36	0.36	0.57	0.60	0.46	0.36	0.27	0.50	0.05	0.52	0.68	0.39	0.98	0.36	0.08	0.53	0.43	0.32	0.52	0.57
t2a-fees	0.29	0.17	0.43	0.72	0.24	0.32	0.27	0.23	0.32	0.38	0.46	0.49	1.00	0.34	0.41	0.38	0.37	0.49	0.57	0.34	0.24	0.45	0.82	0.27	0.34	0.31	0.30	0.38	0.39	0.30	0.22	0.22	0.34	0.05	0.28	0.45	0.28	0.43	0.63	0.07	0.32	0.31	0.22	0.30	0.36
t2a-landing	0.31	0.21	0.30	0.20	0.64	0.24	0.17	0.22	0.22	0.25	0.38	0.32	0.34	1.00	0.27	0.22	0.28	0.30	0.32	0.33	0.24	0.31	0.28	0.80	0.24	0.20	0.25	0.26	0.26	0.20	0.15	0.32	0.24	0.06	0.19	0.27	0.25	0.29	0.18	0.17	0.24	0.20	0.21	0.21	0.24
t2a-privacy-policy	0.40	0.26	0.54	0.30	0.25	0.97	0.29	0.38	0.38	0.63	0.44	0.55	0.41	0.27	1.00	0.33	0.43	0.45	0.70	0.39	0.28	0.54	0.34	0.25	0.97	0.30	0.40	0.40	0.63	0.47	0.44	0.27	0.35	0.05	0.71	0.51	0.42	0.53	0.27	0.09	0.97	0.34	0.37	0.37	0.62
t2a-referrals	0.26	0.14	0.35	0.26	0.19	0.30	0.96	0.20	0.26	0.33	0.33	0.37	0.38	0.22	0.33	1.00	0.26	0.33	0.40	0.28	0.18	0.36	0.31	0.20	0.30	0.97	0.23	0.29	0.33	0.26	0.19	0.18	0.28	0.03	0.27	0.36	0.24	0.35	0.23	0.05	0.30	0.91	0.19	0.25	0.31
t2a-risk-warning	0.27	0.21	0.35	0.21	0.21	0.39	0.20	0.95	0.28	0.49	0.35	0.37	0.37	0.28	0.43	0.26	1.00	0.36	0.56	0.31	0.26	0.36	0.30	0.25	0.40	0.23	0.97	0.32	0.49	0.31	0.23	0.23	0.28	0.05	0.28	0.37	0.26	0.35	0.21	0.07	0.39	0.25	0.90	0.27	0.50
t2a-security	0.33	0.19	0.56	0.30	0.25	0.39	0.26	0.27	0.90	0.47	0.44	0.60	0.49	0.30	0.45	0.33	0.36	1.00	0.58	0.36	0.24	0.57	0.38	0.26	0.40	0.29	0.32	0.93	0.47	0.36	0.27	0.22	0.40	0.04	0.38	0.53	0.30	0.56	0.27	0.07	0.39	0.40	0.26	0.85	0.45
t2a-terms-of-use	0.43	0.27	0.59	0.35	0.27	0.63	0.32	0.48	0.45	0.82	0.55	0.63	0.57	0.32	0.70	0.40	0.56	0.58	1.00	0.43	0.30	0.59	0.39	0.26	0.63	0.33	0.49	0.47	0.82	0.49	0.37	0.28	0.43	0.04	0.54	0.62	0.41	0.59	0.31	0.08	0.62	0.39	0.45	0.43	0.81
t2b-about-us	0.88	0.26	0.42	0.28	0.29	0.39	0.26	0.26	0.33	0.43	0.92	0.43	0.34	0.33	0.39	0.28	0.31	0.36	0.43	1.00	0.34	0.46	0.46	0.38	0.44	0.33	0.35	0.44	0.54	0.33	0.23	0.28	0.37	0.04	0.35	0.44	0.48	0.42	0.24	0.10	0.38	0.30	0.25	0.31	0.40
t2b-buy-crypto	0.27	0.92	0.33	0.20	0.27	0.27	0.16	0.23	0.22	0.30	0.29	0.33	0.24	0.24	0.28	0.18	0.26	0.24	0.30	0.34	1.00	0.35	0.33	0.27	0.31	0.21	0.29	0.30	0.37	0.23	0.17	0.22	0.25	0.06	0.24	0.30	0.24	0.32	0.16	0.08	0.27	0.19	0.21	0.20	0.29
t2b-faq	0.42	0.31	0.99	0.41	0.30	0.53	0.35	0.33	0.55	0.60	0.43	0.99	0.45	0.31	0.54	0.36	0.36	0.57	0.59	0.46	0.35	1.00	0.49	0.32	0.55	0.37	0.38	0.60	0.63	0.46	0.36	0.27	0.50	0.05	0.52	0.68	0.39	0.98	0.36	0.08	0.53	0.43	0.32	0.52	0.57
t2b-fees	0.30	0.19	0.43	0.73	0.25	0.32	0.27	0.24	0.32	0.38	0.35	0.45	0.82	0.28	0.34	0.31	0.30	0.38	0.39	0.46	0.33	0.49	1.00	0.34	0.41	0.38	0.37	0.49	0.56	0.30	0.23	0.22	0.34	0.06	0.29	0.45	0.29	0.43	0.63	0.08	0.32	0.31	0.22	0.30	0.36
t2b-landing	0.32	0.22	0.30	0.21	0.64	0.24	0.18	0.22	0.22	0.26	0.34	0.31	0.27	0.80	0.25	0.20	0.25	0.26	0.26	0.38	0.27	0.32	0.34	1.00	0.27	0.23	0.28	0.30	0.33	0.20	0.16	0.34	0.24	0.08	0.19	0.27	0.26	0.30	0.18	0.16	0.24	0.20	0.21	0.21	0.25
t2b-privacy-policy	0.40	0.26	0.54	0.30	0.25	0.97	0.29	0.38	0.39	0.63	0.40	0.54	0.34	0.24	0.97	0.30	0.40	0.40	0.63	0.44	0.31	0.55	0.41	0.27	1.00	0.33	0.43	0.45	0.70	0.47	0.44	0.27	0.35	0.05	0.71	0.51	0.42	0.53	0.27	0.09	0.97	0.34	0.37	0.37	0.62
t2b-referrals	0.27	0.15	0.35	0.27	0.19	0.30	0.96	0.20	0.26	0.33	0.28	0.36	0.31	0.20	0.30	0.97	0.23	0.29	0.33	0.33	0.21	0.37	0.38	0.23	0.33	1.00	0.26	0.33	0.40	0.26	0.19	0.18	0.28	0.03	0.27	0.36	0.24	0.35	0.23	0.05	0.30	0.91	0.19	0.25	0.31
t2b-risk-warning	0.28	0.23	0.35	0.22	0.22	0.39	0.20	0.95	0.28	0.49	0.31	0.36	0.30	0.25	0.40	0.23	0.97	0.32	0.49	0.35	0.29	0.38	0.37	0.28	0.43	0.26	1.00	0.36	0.56	0.31	0.23	0.24	0.29	0.05	0.28	0.37	0.26	0.35	0.21	0.07	0.39	0.25	0.90	0.27	0.50
t2b-security	0.34	0.21	0.56	0.31	0.25	0.39	0.26	0.28	0.91	0.47	0.37	0.57	0.38	0.26	0.40	0.29	0.32	0.93	0.47	0.44	0.30	0.60	0.49	0.30	0.45	0.33	0.36	1.00	0.58	0.36	0.27	0.22	0.40	0.04	0.38	0.54	0.30	0.56	0.27	0.07	0.39	0.40	0.27	0.85	0.45
t2b-terms-of-use	0.43	0.28	0.60	0.35	0.27	0.63	0.32	0.48	0.45	0.83	0.43	0.60	0.39	0.26	0.63	0.33	0.49	0.47	0.82	0.54	0.37	0.63	0.56	0.33	0.70	0.40	0.56	0.58	1.00	0.49	0.37	0.28	0.43	0.04	0.54	0.63	0.42	0.59	0.31	0.08	0.63	0.39	0.46	0.43	0.81
t3-aml-kyc	0.33	0.22	0.46	0.28	0.21	0.47	0.25	0.30	0.35	0.49	0.33	0.46	0.30	0.20	0.47	0.26	0.31	0.36	0.49	0.33	0.23	0.46	0.30	0.20	0.47	0.26	0.31	0.36	0.49	1.00	0.30	0.35	0.40	0.06	0.54	0.66	0.29	0.46	0.25	0.06	0.47	0.31	0.28	0.33	0.48
t3-cookies-policy	0.24	0.15	0.36	0.21	0.15	0.44	0.19	0.22	0.27	0.37	0.23	0.36	0.22	0.15	0.44	0.19	0.23	0.27	0.37	0.23	0.17	0.36	0.23	0.16	0.44	0.19	0.23	0.27	0.37	0.30	1.00	0.22	0.25	0.04	0.48	0.35	0.24	0.36	0.19	0.05	0.44	0.23	0.21	0.26	0.37
t3-landing	0.26	0.20	0.27	0.17	0.30	0.27	0.16	0.20	0.19	0.28	0.28	0.27	0.22	0.32	0.27	0.18	0.23	0.22	0.28	0.28	0.22	0.27	0.22	0.34	0.27	0.18	0.24	0.22	0.28	0.35	0.22	1.00	0.25	0.10	0.31	0.30	0.26	0.26	0.14	0.12	0.26	0.17	0.19	0.17	0.27
t3-news	0.37	0.23	0.49	0.31	0.25	0.35	0.27	0.26	0.39	0.43	0.37	0.50	0.34	0.24	0.35	0.28	0.28	0.40	0.43	0.37	0.25	0.50	0.34	0.24	0.35	0.28	0.29	0.40	0.43	0.40	0.25	0.25	1.00	0.05	0.36	0.59	0.28	0.49	0.27	0.07	0.35	0.33	0.25	0.37	0.41
t3-nft	0.03	0.05	0.05	0.03	0.06	0.04	0.02	0.03	0.03	0.04	0.04	0.05	0.05	0.06	0.05	0.03	0.05	0.04	0.04	0.04	0.06	0.05	0.06	0.08	0.05	0.03	0.05	0.04	0.04	0.06	0.04	0.10	0.05	1.00	0.06	0.06	0.03	0.05	0.04	0.04	0.05	0.03	0.04	0.04	0.04
t3-privacy-notice	0.36	0.22	0.52	0.28	0.21	0.71	0.27	0.28	0.38	0.54	0.35	0.52	0.28	0.19	0.71	0.27	0.28	0.38	0.54	0.35	0.24	0.52	0.29	0.19	0.71	0.27	0.28	0.38	0.54	0.54	0.48	0.31	0.36	0.06	1.00	0.54	0.37	0.52	0.25	0.07	0.71	0.33	0.27	0.36	0.53
t3-terms-of-service	0.43	0.27	0.68	0.42	0.28	0.51	0.35	0.34	0.52	0.63	0.44	0.68	0.45	0.27	0.51	0.36	0.37	0.53	0.62	0.44	0.30	0.68	0.45	0.27	0.51	0.36	0.37	0.54	0.63	0.66	0.35	0.30	0.59	0.06	0.54	1.00	0.36	0.67	0.36	0.07	0.51	0.44	0.33	0.49	0.62
t4-about-us	0.52	0.25	0.40	0.29	0.26	0.42	0.25	0.26	0.31	0.42	0.48	0.39	0.28	0.25	0.42	0.24	0.26	0.30	0.41	0.48	0.24	0.39	0.29	0.26	0.42	0.24	0.26	0.30	0.42	0.29	0.24	0.26	0.28	0.03	0.37	0.36	1.00	0.45	0.51	0.14	0.48	0.38	0.39	0.45	0.51
t4-faq	0.42	0.30	0.99	0.41	0.30	0.53	0.35	0.33	0.55	0.59	0.42	0.98	0.43	0.29	0.53	0.35	0.35	0.56	0.59	0.42	0.32	0.98	0.43	0.30	0.53	0.35	0.35	0.56	0.59	0.46	0.36	0.26	0.49	0.05	0.52	0.67	0.45	1.00	0.45	0.10	0.55	0.46	0.37	0.58	0.61
t4-fees	0.26	0.16	0.37	0.67	0.22	0.27	0.24	0.22	0.29	0.31	0.24	0.36	0.63	0.18	0.27	0.23	0.21	0.27	0.31	0.24	0.16	0.36	0.63	0.18	0.27	0.23	0.21	0.27	0.31	0.25	0.19	0.14	0.27	0.04	0.25	0.36	0.51	0.45	1.00	0.14	0.37	0.44	0.43	0.54	0.47
t4-landing	0.10	0.08	0.08	0.07	0.15	0.09	0.05	0.07	0.07	0.08	0.10	0.08	0.07	0.17	0.09	0.05	0.07	0.07	0.08	0.10	0.08	0.08	0.08	0.16	0.09	0.05	0.07	0.07	0.08	0.06	0.05	0.12	0.07	0.04	0.07	0.07	0.14	0.10	0.14	1.00	0.11	0.09	0.11	0.11	0.11
t4-privacy-policy	0.40	0.26	0.53	0.30	0.25	0.97	0.29	0.39	0.39	0.63	0.39	0.53	0.32	0.24	0.97	0.30	0.39	0.39	0.62	0.38	0.27	0.53	0.32	0.24	0.97	0.30	0.39	0.39	0.63	0.47	0.44	0.26	0.35	0.05	0.71	0.51	0.48	0.55	0.37	0.11	1.00	0.38	0.42	0.43	0.69
t4-referrals	0.31	0.18	0.43	0.31	0.22	0.34	0.92	0.25	0.41	0.39	0.30	0.43	0.31	0.20	0.34	0.91	0.25	0.40	0.39	0.30	0.19	0.43	0.31	0.20	0.34	0.91	0.25	0.40	0.39	0.31	0.23	0.17	0.33	0.03	0.33	0.44	0.38	0.46	0.44	0.09	0.38	1.00	0.32	0.49	0.44
t4-risk-warning	0.26	0.21	0.32	0.22	0.21	0.37	0.19	0.92	0.27	0.46	0.25	0.32	0.22	0.21	0.37	0.19	0.90	0.26	0.45	0.25	0.21	0.32	0.22	0.21	0.37	0.19	0.90	0.27	0.46	0.28	0.21	0.19	0.25	0.04	0.27	0.33	0.39	0.37	0.43	0.11	0.42	0.32	1.00	0.41	0.54
t4-security	0.32	0.19	0.53	0.30	0.24	0.37	0.25	0.27	0.88	0.43	0.31	0.52	0.30	0.21	0.37	0.25	0.27	0.85	0.43	0.31	0.20	0.52	0.30	0.21	0.37	0.25	0.27	0.85	0.43	0.33	0.26	0.17	0.37	0.04	0.36	0.49	0.45	0.58	0.54	0.11	0.43	0.49	0.41	1.00	0.53
t4-terms-of-use	0.41	0.27	0.57	0.34	0.25	0.62	0.31	0.49	0.44	0.81	0.40	0.57	0.36	0.24	0.62	0.31	0.50	0.45	0.81	0.40	0.29	0.57	0.36	0.25	0.62	0.31	0.50	0.45	0.81	0.48	0.37	0.27	0.41	0.04	0.53	0.62	0.51	0.61	0.47	0.11	0.69	0.44	0.54	0.53	1.00

Figure 7.9: Pairwise content similarities for representative websites.

8

Discussion

8.1. Discussion of results

This section discusses the implications of the results and explores their significance in the context of the research questions.

8.1.1. Lifecycle

From the results, it became clear that the websites are likely to have a life cycle of exactly one year. Auto-renewal of domains is very common for websites, but requires payment. The finding that websites have a lifespan of one year suggests that these websites have been a one-time investment. This can be explained by either an inability to pay or by unwillingness to pay. The price of registering a domain for a year partially depends on the top level domain. A *.com* domain can be bought for a few euros per year. Inability to pay therefore does not only refer to the access to financial means, but also to payment in a way that cannot be traced back to the creator. A possible explanation for unwillingness to pay is limited success of a website. If the website has not reached victims in the first year, then a secondary financial investment, however small, might not be worth it.

The one-year lifespan of the websites also means that no conclusions about the age of the templates could be drawn from the heatmap of domain creation dates in figure 7.1. The lower bound of the template's age can be set to one year, but the exact age is difficult to determine. As said before, the websites with a current lifespan of over a year do not necessarily indicate that the template has the same minimum age.

Internet articles about the specific scams can be used to determine the minimum age of the templates. Websites and videos that warn for scams have made reports about websites with template 1 and 2a that were registered in April 2023 [32–35]. This indicates that these templates are at least 19 months old.

Furthermore, the one-year lifespan means that drawing conclusions about the popularity of the template over time is more difficult. From the heatmap it seems that templates 1 and 2 show a decrease in popularity over time, meaning that less new domains get created over time. However, it is difficult to determine whether this is representative for the full scam operation, since there are no data points from the first appearances of these templates available anymore and since it is unclear what the exact delay of Google's website indexing is.

Overall, a lifespan of exactly one year means that these websites are undisturbed during their scamming operation and go offline at the moment the website owner expects the website to go offline. This means that during this one year period, no takedowns of these websites have been successfully executed. This is either due to these websites being hosted by bulletproof hosting providers, who actively protect the

websites, or simply due to a lack of enforcement action; no authority has made the effort to initiate its takedown.

8.1.2. Intra-template similarity

Overall, the results of the intra-template similarity analysis show that these websites use the same structure and hardly deviate from the template. For the textual content, the similarities are overall slightly lower, but the high frequency of mean values above 0.7 shows some evidence that the template and textual content are offered as a package. Especially the similarity values for which the domain names have been replaced by placeholders show that many websites only differ in website name.

Additional evidence for this hypothesis was found online. The Google search query *"crypto trading with a leading exchange" scam* suggests a post on the Breachforums website. This can be seen in figure 8.1 It seems that one of the templates was being discussed on or offered through breachforums.st. The presence of the distinct phrase proves that the post is related to the templates discussed in this study. Breachforums is a crime forum, which has been seized multiple times. This post provides additional evidence for the hypothesis that the template has been offered together with the textual content and that the malicious content is not added after acquiring the template. This underlines the importance of finding the listings of such templates in criminal investigations: in the case a listing exists, it can prove malicious intent of the template provider, as well as of the website owners.

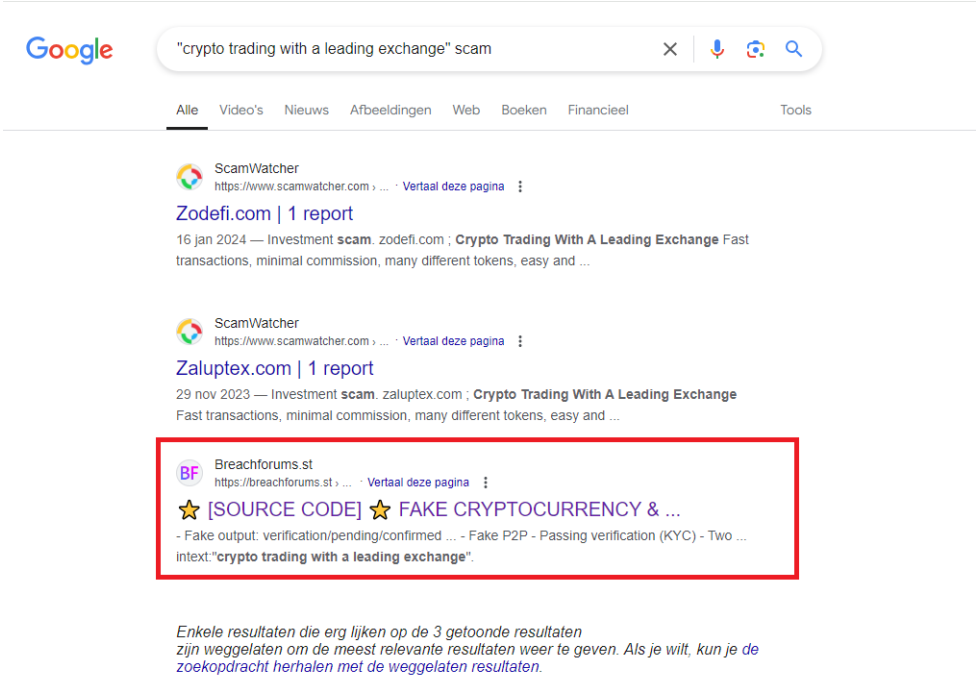


Figure 8.1: A screenshot of the Breachforums suggestion.

The *fees*, *terms of use*, and *aml-kyc* pages had the lowest textual content similarity. Of these pages, the *terms of use* of template 1 has the least unity, meaning that the relatively high standard deviations cannot be explained by a separation into two groups or by a few extreme outliers. This can indicate that this webpage is personalized the most. A logical explanation for this can be that when a victim is asking why they are not getting their money back, the scammers can refer to the terms of use to ensure them that the website owners are not in the wrong. This could become the buildup to an advanced fee scam. By personalizing the *terms of use* page, the scammers can personalize the way they perform an advanced fee scam. Text box 8.1.2 illustrates a difference within a template. It shows the first two paragraphs of the terms of use pages of template 2.

Textbox 8.1.2: Terms of Use comparison (ID= 2-5 and 2-20)

1. TERMS OF USE

1.1 Acceptance of the these terms of use. By accessing, reading, and making use of this Website and the Services, you are deemed to have reviewed, understood and accepted, on your own behalf and on behalf of any person on whose behalf you may be acting, these Terms of Use and agreed with the Company to be bound hereunder. For the purposes of these Terms of Use, “person” means any natural person, corporation, partnership, joint venture or any other incorporated or unincorporated entity, whether acting as an individual, fiduciary, or in any other capacity.If you do not wish to be bound by these Terms of Use, you may not use this Website or any of the Services.

1.2 Changes of terms of use and website. The Company reserves the right, exercisable at any time at its sole discretion, to add to or remove, modify or otherwise change any part of these Terms of Use. Changes will be effective immediately at such time as the Terms of Use are posted on this Website. You should check the Terms of Use for changes by checking the date this page was last updated. If any change to the Terms of Use is not acceptable to you, you must discontinue your use of this Website and the Services immediately. Your continued use of this Website or any of the Services after any changes to the Terms and Conditions will constitute your unqualified acceptance of the changes. The Company may terminate, change, suspend or discontinue any aspect of this Website or the Services at any time without notice. Without limiting the generality of the foregoing, the Company may change the availability of any features, institute new, or amend existing fees or charges for the use of the Website, the Services, or any features included in the Website or the Services, add, remove, modify or otherwise change any content on this Website, and impose limits on certain features or restrict access to parts or all of this Website. The Company reserves the right, but not the duty, to correct any errors or omissions in any portion of this Website at any time and without notice.

1. TERMS OF USE

1.1 By accessing and utilizing this website and its services, you confirm that you have comprehensively read, understood, and consented to be bound by these terms of use. This agreement encompasses both you and any entity you might represent. Should you find these terms unacceptable, you are advised not to access or utilize the website and its associated services.

1.2 The company reserves the exclusive right to modify these terms of use, the website’s content, and any associated services at any given time. Such alterations become effective immediately upon their publication on the website. It’s incumbent upon you to periodically review this page to remain informed about any updates. Should you find any modifications objectionable, your recourse is to cease using the website and its services. Persisting in the use of the website post-modifications signifies your acceptance of those changes. Additionally, the company may, without prior notification, alter, suspend, or terminate any aspect of the website or its services. This includes, but isn’t limited to, feature adjustments, the introduction of new charges, content updates, and the imposition of access restrictions to specific sections of the website. The company also retains the right to rectify any inaccuracies or oversights on the website without prior notice.

8.1.3. Inter-template similarity

The inter-template similarities show that content is shared between templates and that the landing pages are the most unique pages. A logical explanation is that the landing page is seen first and tries to leave a good impression. Therefore, it is the most complex webpage of the website. If assumed that the domain owner is not necessarily the template creator, this can be reasoned from the template creator’s perspective: the landing page is the business card of the template itself. In other words, potential scammers should be persuaded to buy the template, which is easier with a distinctive landing page.

By analyzing figures 7.6 and 7.8 together, it can be seen that the structural similarities within a cluster are much higher than the structural similarities between the clusters. Although this is logical, it does prove that fingerprinting the individual templates based on the HTML structure of a website is possible. This demonstrates that new instances of the template can be detected through fingerprinting. This can be achieved by scanning newly registered domains for the known template. Lists of newly created domains can be found through several, mostly paid, services[1], or through SSL certificate transparency logs.

Similarly, new templates that share textual content with known templates can be discovered by fingerprinting the textual content of the known template. Monitoring newly created websites and sending Notice and Takedown requests quickly after a website appearing is the most effective way of ensuring that scam domain will not have any victims and is therefore an important measure in the battle against online scamming. By taking down websites that share the textual content or structure with known scam templates, the current modus operandi is disrupted. This disruption elevates the level of effort needed in order to create a new scam website. Depending on the scam, this increase of effort needed either lowers the scammers profit, or makes the whole scam operation economically unprofitable.

8.1.4. Cloudflare

An integral part of these scams was the use of the Cloudflare Turnstile page. Cloudflare provides websites with anti-bot protection, which prevents the websites from getting DDoSsed. It hides the actual IP address of a website and serves as a proxy for these domains. As said before, this also makes automatic scraping more difficult and limits the ability of archiving tools to take a snapshot of the content. The positive impact that Cloudflare has on the safety of the internet is not being questioned. Especially in an age of generative AI, in which technologies rely upon large quantities of scraped data, people should seriously consider protecting their websites against scraping. However, this study's use case illustrates that Cloudflare can also be an important part of a scammers deviant security strategy. It provides an extra layer of anonymity and the lack of archived data makes it more difficult to prove what a website looked like at the time of a scam. These Cloudflare services are all part of a free plan, which makes that there is no payment trail needed.

Arguably the most important deviant security asset that Cloudflare offers is its ability to obfuscate the real geolocation of the server hosting the scam domain, by acting as a proxy. An information deficit regarding the geolocation of a website's server can be reason enough for law enforcement not to start an investigation. Scammers that use a proxy add an extra hop to the path towards prosecution by obfuscating the feasibility of attribution, as well as the geographical relevance to the law enforcement agencies. If law enforcement does not know with certainty whether the suspects are in their jurisdiction or in the jurisdiction of another cooperating nation, then it is unclear what the chances of a successful prosecution are. Logically, a lack of information about the geolocation also results into a lack of takedown efforts of law enforcement agencies, because there is no clear reason why a certain country's law enforcement should spend valuable resources instead of any other country. The scam operations are currently profiting from this created accountability void.

Cloudflare declares to cooperate with law enforcement requests, however they state that "[u]nless there is an emergency, Cloudflare requires valid legal process such as a subpoena before providing [customer] information" [36]. This means that Cloudflare will comply with providing the necessary information for ongoing investigations, but will not provide information a voluntary basis. This has as a result that the information Cloudflare holds cannot be used in the early stages in which law enforcement is still deciding whether they have enough cause and prospect to start an investigation. In the case that law enforcement does not want to start an investigation, but rather chooses to disrupt the scam operation, then they have to rely on relaying information to Cloudflare and trusting them to deal with abuse reports. This illustrates the position of power that Cloudflare has as link in the investigation chain.

[1] Examples include https://newly-registered-domains.whoisxmlapi.com/ and https://research.domaintools.com/phisheye/

8.1.5. Template change

On June 6 2024, an overall template change for template 1 was noticed. A *buy-crypto* page and a header with *NFT - Soon* were added. This change went into effect simultaneously across most websites with this template. This suggests that this template is updated from a centralized location. It does not necessarily mean that all websites are owned by the same entity, but it does mean that all updated websites have a common supplier. This shows that monitoring such clusters of websites can provide additional information about the scamming operation as a whole.

8.2. Discussion of recommendations

This section outlines the recommendations that should guide law enforcement in their investigative decisions.

8.2.1. Balance between law enforcement efforts and scamming efforts

The scam websites analyzed in this study illustrate that the set-up of scam operations require minimal effort. Cloudflare provides an important layer of anonymity and domain registrant data can be hidden for privacy reasons. Assuming the website template is offered as a product, in order to be operational, the scammer must register a domain, buy and host the website template, and activate the Cloudflare services. These are all actions that only need to be performed once, so scammers can follow a 'fire and forget' approach. Although a working scam website does not yet guarantee a scam's success, it does show that initial efforts in order to launch a website, or a multitude of websites, is minimal.

As described in subsection 8.1.4, the level of effort required to prosecute the entity behind a scam website is currently so substantial that legal action is often not pursued. The Notice and Takedown route is more feasible: it achieves the same goal of protecting citizens from getting scammed and requires no formal criminal proceedings. Currently, the lack of a streamlined NTD pipeline prevents law enforcement from effectively countering scam websites. By automating a significant part of the NTD process in a transparent way, while maintaining human verification as an essential component, the resources required for each individual takedown will decrease. This will likely result in a more frequent pursuit of the Notice and Takedown procedure. The automation should include suggesting the takedown of websites that fit the structural or textual fingerprint of previously discovered scam templates.

From a cost perspective, deploying a scam website is comparable to a normal website. The domain registration and hosting need to be paid upfront, often either monthly or yearly. Website templates are cheap, but a customized websites can be expensive. The Cloudflare Turnstile services are free, regardless of traffic volume.

In contrast, tools and datasets that collect and combine website information are often paid services, especially when consulting is needed on a large scale. Examples include registrant data lookup tools or inventories of newly created domains. Some services offer licenses to law enforcement, but given the value these tools add to existing data, they are typically available only as paid services. Therefore, an efficient Notice and Takedown pipeline requires upfront investments, either by acquiring tooling developed by private companies or by creating similar tools in-house.

A country is primarily responsible for the safety of its own citizens, but citizens can fall victim to foreign websites or websites with an unknown geolocation. The absence of geolocation information currently creates an accountability void, therefore it is important not to limit the NTD process to domestic websites and to look beyond borders.

By investing effort and money into the automation of the Notice and Takedown process, the effort needed by law enforcement to take down a scam operation will decrease and will simultaneously increase the effort and financial means needed by scammers to keep their scam operational.

8.2.2. Clustering techniques

Law enforcement can use a clustering analysis, as performed in this study, to inform their decision on how to focus their investigative efforts. Clustering can aid in providing intuition of which domains can, or cannot, be attributed to the same entity. On the one hand, the websites that did not belong to the

core cluster can pose investigative opportunities, since they show that they do not follow the standard methods of this cluster. The creators of these outliers might have made mistakes in their attempt to improve upon the template. On the other hand, these outliers might be considered low hanging fruit. If the goal is disrupting the larger operation, then efforts should be aimed at taking down the core cluster(s) of a template.

Another problem that can benefit from clustering is the problem of police reports not ending up on the same paper pile when the domain name is different, but the template is the same. Website copies can be identified more easily by systematically collecting screenshots or HTML content of fraudulent websites and regularly aiming to cluster police reports accordingly to detect scam operations worth investigating.

8.2.3. Scraping techniques

This study contained an empirical evaluation of the feasibility of investigating scam clusters at scale. Due to the Cloudflare Turnstile page, the usual scraping tools, such as Selenium, did not work. This caused the use of a scraping method that needs some human supervision, as described in section 5.7. A limiting difference between the two methods is the inability to specify explicit waits with PyAutoGUI. Selenium allows waiting until a certain element is loaded, which provides certainty that the page is loaded correctly. The working of PyAutoGUI relies on implicit waits, such as *wait for 4 seconds*. Since this is more dependent on a reliable internet connection, the waits need to be assigned with generous margins in order to ensure a high likelihood of scraping a fully loaded page. This logically leads to longer overall scraping times, which will inevitably have a negative influence on the feasibility of scraping at large scale.

The current method using PyAutoGUI works well for websites with identical layout. The websites are mapped into an opcode file once per template. It uses clicking through the website and is therefore dependent on this identical layout. A way to solve this is to directly input the subdomain URL into the URL bar. This method can better deal with varying content, but is dependent on knowing the subdomain URLs. This way of scraping exhibits behavior that is less human-like, which increases the likelihood of being flagged as non-human.

Depending on the needs of law enforcement, scraping can be used in different ways. It can be used to collect evidence of what the contents of a certain website look like at certain times, which can be used in court cases. Secondly, scraping can be used to collect input data for a fingerprint based scan for known templates. New instances of known templates can be discovered by scraping and analyzing newly registered domains, or only the anomalies of a URL-based anomaly detection if a smaller scope is needed.

8.2.4. Reverse registrant lookup

Many of these scam websites hide their registrant information, which is personal information about the owner of the website. This is mostly out of anonymity reasons. However, some domains did reveal a registrant name. By leveraging reverse registrant lookup tooling[2], it was possible to connect multiple templates to one registrant. This leads to the hypothesis that the different templates share the same source. Although this method is not fully conclusive – since multiple people can register a website under the same name – the combination of a unique enough registrant name, registrar and domain name does lead to a intuition of which websites are hosted by the same entity. Reverse registrant lookups can therefore be utilized by law enforcement to discover additional scam domains and to tie multiple websites to the same entity. However, this can only be done in the rare cases in which the registrant name is not shielded.

[2] https://www.whoxy.com/

9

Limitations and Future Works

This chapter discusses the limitations of the study, and the potential for further research of this topic.

9.1. Google search API

A limiting factor of the dataset quality is the fact that the Google Search API is a black box. This means that it is unknown which websites are actively not indexed based on Google's anti-scam efforts [17]. Furthermore, it is unknown how long it takes before these websites appear in Google Search suggestions. Thus, websites that go offline before they have been indexed cannot be captured. This causes the dataset to suffer from survivorship bias. It is difficult to fully determine the representativeness of the templates in the dataset for the whole cluster of that template, as well as the representativeness of these templates for the whole scam landscape.
In future research, a different dataset can be created by fingerprinting existing scam templates and consulting inventories of newly created domains to find new websites containing these templates. This additional step in the methodology can overcome the black-box hurdle and survivorship bias and could more accurately assess the popularity of a template over time. However, this methodology will not capture any websites of which the content has been altered long after their creation. Therefore, it would best function as an addition to the current methodology, rather than as a replacement of it.

9.2. Content originality

The templates reviewed in this study contain quite some textual content that is unique for scam websites, yet shared across many different templates. Alternatively, the most frequently occurring *about us* text has been copied from a description of Kraken: a legitimate cryptocurrency exchange.

Both these cases underline the lack of originality regarding the textual content. The websites in this research have highlighted a 'low-investment' modus operandi, which is accompanied by a lack of originality in the textual content. However, the general idea that customization of the template costs time and effort might already be outdated. With the arrival of the new generative AI possibilities, it will be easier to customize the templates and evade fingerprinting based detection methods. Therefore, future research can focus on creating robust classifiers: the detections must not only rely on the structural and textual content of a website, but also on patterns in the available metadata, such as the domain registrar and SSL certificate data.

9.3. Effectiveness of scam operation

It is difficult to determine how effective the scams shown in this study are. The Pamonex website – which was a scam of template 2 that launched in April 2023 – had some victims, but the scam was quickly unveiled. Websites that collect scam reports, such as Chainabuse, did not provide evidence of recent victims. However, the persistence of the creation of new domains with the scam templates does indicate that the scammers have not completely given up on the effectivity of the template. This lack of victims

causes the questioning of the effectiveness of the scam operation, for which a multitude of options are plausible. One possibility is that the scam has not yet been effective, but that the scammers are hopeful that it will be effective in the future. Another possibility is that the scam is effective, but that no reports are available online. This can be due to the victims not realizing that they have been scammed, them consciously not reporting the scam, or because the victims have reported it, but the reports have not been made publicly available. A third possibility works under the assumption that the template is a product that can be bought by other scammers. If the template seller presents the template as a way to make buyers rich without having to do anything, whereas in reality the websites are not getting any visitors, then it might actually be the template buyers who are getting scammed. Which of the scenarios holds the truth cannot be answered in this study since there is insufficient evidence to support any of the scenarios.

10

Conclusion

This study attempted to capture the tactics, techniques, and procedures of the owners and creators of cryptocurrency investment scam websites. By monitoring a cluster of 430 scam websites for 40 days, we discovered that the median lifespan of these websites was exactly one year, which was caused by the expiration of the domain registration. Websites within the same template cluster hardly contain differences and if changes occurred, then this was most likely done on pages that could be leveraged as source to explain why the victim should pay extra fees or why they cannot withdraw their money.

The free services offered by Cloudflare to provide security and anonymity for any website and the obfuscation of registrant data are two methods used by the scammers to create an extra hop on the route towards attribution. Additionally, using Cloudflare prevents the website contents from being archived. The obfuscation of geolocation data causes an accountability void. The lack of clarity about whose jurisdiction the website's hoster is located in prevents law enforcement from taking decisive action.

Although we found no clear indicators that would suggest the feasibility of prosecution, the results show that utilizing clustering techniques might help with attributing websites to different owners, or even the same owner in the case that registrant data was not hidden. Registrant data revealed that a user can have multiple domains with different templates registered.

Overall, this research contains evidence to support the theory that a template has one provider and is bought by multiple entities, and evidence for the theory that one entity is the sole owner of the template with the existence of some copycats. Evidence supporting the first theory is the discussion of the website template on a black-hat crime forum, the presence of subtle changes within the templates, and the ongoing addition of new domains. Within this theory, clusters showcasing a similar modus operandi can be explained by individual entities that launch a high number of new websites.

Evidence supporting the second theory is the large core clusters that follow a similar modus operandi. Another observation that supports this theory was the template change of template 1. This suggested that the websites are centrally updated, rather than being self-hosted. The observation that some websites did not get updated supports the copycat theory: these websites share the same content, but the template is not provided through the same channels.

This study highlighted the difficulties that law enforcement face in the course of actions against scam websites and advises investments in a more structured Notice and Takedown pipeline with attention to detection of recurrent content. The similarities found in this study display the potential of detecting websites using the fingerprint of known scams to find both known and novel templates.

Additionally, law enforcement should be aware of the multiple different paper piles scam reports might end up on and should actively try to combine paper piles to increase the likelihood that a case can be identified as worth investigating.

Although the chances of convicting the owners of scam remain low, it is encouraging to see that such a large online community actively works to expose and combat these fraudulent websites.

References

[1] S. Nakamoto, "Bitcoin: A peer-to-peer electronic cash system," Oct. 2008. [Online]. Available: https://bitcoin.org/bitcoin.pdf.

[2] Blockchain.com. "Total number of transactions." Accessed: November 26, 2024., [Online]. Available: https://www.blockchain.com/explorer/charts/n-transactions-total.

[3] CoinMarketCap. "Bitcoin." Accessed: April 28, 2024., [Online]. Available: https://coinmarketcap.com/currencies/bitcoin/.

[4] M. Bartoletti, S. Lande, A. Loddo, L. Pompianu, and S. Serusi, "Cryptocurrency scams: Analysis and perspectives," *IEEE Access*, vol. 9, pp. 148 353–148 373, 2021. DOI: 10.1109/ACCESS.2021.3123894.

[5] L. Nizzoli, S. Tardelli, M. Avvenuti, S. Cresci, M. Tesconi, and E. Ferrara, "Charting the landscape of online cryptocurrency manipulation," *IEEE Access*, vol. 8, pp. 113 230–113 245, 2020. DOI: 10.1109/access.2020.3003370.

[6] Chainalysis, *"The 2024 Crypto Crime report"*, Feb. 2024. [Online]. Available: https://go.chainalysis.com/crypto-crime-2024.html.

[7] European Commission, "Regulation (eu) 2022/2065 of the european parliament and of the council of 19 october 2022 on a single market for digital services and amending directive 2000/31/ec (digital services act)," *Official Journal of the European Union*, vol. 65, 2022. [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32022R2065.

[8] A. Kuczerawy, "From 'Notice and Takedown' to 'Notice and Stay Down': Risks and Safeguards for Freedom of Expression," in *Oxford Handbook of Online Intermediary Liability*, Oxford University Press, May 2020. DOI: 10.1093/oxfordhb/9780198837138.013.27.

[9] *Wetboek van strafrecht*, Oct. 2024. [Online]. Available: https://wetten.overheid.nl/jci1.3:c:BWBR0001854&boek=Tweede&titeldeel=XXV&artikel=326&z=2024-10-01&g=2024-10-01.

[10] C. Cross, "Romance baiting, cryptorom and 'pig butchering': An evolutionary step in romance fraud," *Current Issues in Criminal Justice*, pp. 1–13, 2023. DOI: 10.1080/10345329.2023.2248670.

[11] D. Bianco. "The pyramid of pain." Accessed: April 26, 2024., [Online]. Available: https://detect-respond.blogspot.com/2013/03/the-pyramid-of-pain.html.

[12] CoinMarketCap. "Historical snapshot - 22 october 2017." Accessed: April 26, 2024., [Online]. Available: https://coinmarketcap.com/historical/20171022/.

[13] CoinMarketCap. "Bitconnect." Accessed: April 26, 2024., [Online]. Available: https://coinmarketcap.com/currencies/bitconnect.

[14] Office of Public Affairs of the U.S. Department of Justice, *Crypto fraud victims receive over $17 million in restitution from bitconnect scheme*, Jan. 2023. [Online]. Available: https://www.justice.gov/opa/pr/crypto-fraud-victims-receive-over-17-million-restitution-bitconnect-scheme.

[15] T. Moore, J. Han, and R. Clayton, "The postmodern ponzi scheme: Empirical analysis of high-yield investment programs," in *Financial Cryptography and Data Security*, A. D. Keromytis, Ed., Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 41–56. DOI: 978-3-642-32946-3.

[16] P. Xia *et al.*, "Characterizing cryptocurrency exchange scams," *Computers & Security*, vol. 98, p. 101 993, 2020. DOI: https://doi.org/10.1016/j.cose.2020.101993.

[17] E. Tucker. "New ways we're tackling spammy, low-quality content on search." Accessed: November 6, 2024., [Online]. Available: https://blog.google/products/search/google-search-update-march-2024/.

[18] L.-A. Fenge and S. Lee, “Understanding the Risks of Financial Scams as Part of Elder Abuse Prevention,” *The British Journal of Social Work*, vol. 48, no. 4, pp. 906–923, Jul. 2018. DOI: 10.1093/bjsw/bcy037.

[19] J. Drew and T. Moore, “Optimized combined-clustering methods for finding replicated criminal websites,” *EURASIP Journal on Information Security*, vol. 2014, Dec. 2014. DOI: 10.1186/s13635-014-0014-4.

[20] J. Drew and T. Moore, “Automatic identification of replicated criminal websites using combined clustering,” in *2014 IEEE Security and Privacy Workshops*, 2014, pp. 116–123. DOI: 10.1109/SPW.2014.26.

[21] K. Toyoda, T. Ohtsuki, and P. T. Mathiopoulos, “Identification of high yielding investment programs in bitcoin via transactions pattern analysis,” in *GLOBECOM 2017 - 2017 IEEE Global Communications Conference*, 2017, pp. 1–6. DOI: 10.1109/GLOCOM.2017.8254420.

[22] K. Toyoda, P. Takis Mathiopoulos, and T. Ohtsuki, “A novel methodology for hyip operators’ bitcoin addresses identification,” *IEEE Access*, vol. 7, pp. 74 835–74 848, 2019. DOI: 10.1109/ACCESS.2019.2921087.

[23] W. Chen, Z. Zheng, J. Cui, E. Ngai, P. Zheng, and Y. Zhou, “Detecting ponzi schemes on ethereum: Towards healthier blockchain technology,” in *Proceedings of the 2018 World Wide Web Conference*, International World Wide Web Conferences Steering Committee, 2018, pp. 1409–1418. DOI: 10.1145/3178876.3186046.

[24] G. A. Siu, A. Hutchings, M. Vasek, and T. Moore, ““invest in crypto!”: An analysis of investment scam advertisements found in bitcointalk,” in *2022 APWG Symposium on Electronic Crime Research (eCrime)*, 2022, pp. 1–12. DOI: 10.1109/eCrime57793.2022.10142100.

[25] “Detect malicious websites by building a neural network to capture global and local features of websites,” *Computers & Security*, vol. 137, p. 103 641, 2024. DOI: https://doi.org/10.1016/j.cose.2023.103641.

[26] E. Van De Sandt, “Deviant security: The technical computer security practices of cyber criminals,” Ph.D. dissertation, University of Bristol, 2019.

[27] Harisudhan, *Phishing and legitimate urls*, Accessed: November 7, 2024. [Online]. Available: https://www.kaggle.com/datasets/harisudhan411/phishing-and-legitimate-urls.

[28] B. Wolters, M. Guerreiro, and A. Martinetti. “Cloudflare is free of captchas; turnstile is free for everyone.”, [Online]. Available: https://blog.cloudflare.com/turnstile-ga/.

[29] Cloudflare. “Cloudflare Turnstile.” Accessed: December 2, 2024., [Online]. Available: https://www.cloudflare.com/nl-nl/products/turnstile.

[30] T. Gowda and C. A. Mattmann, “Clustering web pages based on structure and style similarity (application paper),” in *2016 IEEE 17th International Conference on Information Reuse and Integration (IRI)*, 2016, pp. 175–180. DOI: 10.1109/IRI.2016.30.

[31] P. E. Black, “Ratcliff/obershelp pattern recognition,” National Institute of Standards and Technology, Tech. Rep., 2021. [Online]. Available: https://xlinux.nist.gov/dads/HTML/ratcliffObershelp.html.

[32] CryptoNews. “Pamonex raises concerns of being an advanced scam.” Accessed: November 6, 2024., [Online]. Available: https://www.binance.com/en/square/post/803697.

[33] Harold. “Is pamonex exchange a scam or legit?” Accessed: November 6, 2024., [Online]. Available: https://coincu.com/249494-is-pamonex-exchange-a-scam-or-legit/.

[34] J. Liles. “Notexer scam, quickly explained.” Accessed: November 6, 2024., [Online]. Available: https://www.youtube.com/watch?v=hVp-SLuHvoE.

[35] Trustpilot. “Trustpilot - pamonex.” Accessed: November 6, 2024., [Online]. Available: https://www.trustpilot.com/review/pamonex.com.

[36] Cloudflare. “Law enforcement.” Accessed: November 7, 2024., [Online]. Available: https://www.cloudflare.com/trust-hub/law-enforcement/.

A

Dataset overview

Table A.1: Dataset overview

Domain Name	Template	Scraping-ID	Creation Date	Offline Date
24bitswap.ru	1	1-0	24/07/2023	
adevour.com	1	1-1	30/12/2023	
adiswap.com	2	2-0	04/04/2024	
aerorium.com	1	1-2	18/12/2023	
airdaos.com	2	2-1	16/02/2024	
alivlab.com	1	1-3	13/08/2024	
allbtcoin.com	2	2-2	16/09/2023	17/09/2024
alllbtc.com	2	2-3	10/10/2023	10/10/2024
alt-ave.com	3	3-0	22/02/2024	
altowaves.com	1	1-4	31/07/2024	
altowax.com	1	1-5	09/06/2024	
alturaex.com	2		02/09/2023	03/09/2024
altzex.com	1	1-6	22/08/2024	
ameubit.com	1	1-7	30/09/2023	30/09/2024
arbitexo.com	1	1-8	05/12/2023	
arbitsrum.com	2		15/10/2023	
arnaultex.com	1	1-9	14/10/2023	
auroexs.com	1	1-10	04/09/2023	05/09/2024
axutex.com	1	1-11	25/11/2023	
bananses.com	2	2-4	30/10/2023	
basewalletcoins.com	3	3-1	25/06/2024	
beastox.com	1	1-12	30/01/2024	
beastxdefi.com	4	4-0	28/04/2024	
beastxmax.com	1	1-13	03/01/2024	
beastxswap.com	1	1-14	21/02/2024	
beastxwallet.com	1	1-15	21/02/2024	
bebspace.com	1	1-16	06/03/2024	
beconax.com	1	1-17	14/08/2024	
beemorecoins.com	1	1-18	18/07/2024	
besbufex.com	1	1-19	03/10/2023	04/10/2024
betobit.net	1		29/08/2023	
bexdex.com	1	1-20	16/07/2024	
binansium.com	2	2-5	22/10/2023	
binfells.com	2	2-6	05/07/2024	
binplace.com	3	3-2	27/06/2024	
bintendo.com	1	1-21	21/07/2024	
bit-blaze.com	3	3-3	22/02/2024	
bit-blitz.com	3	3-4	22/02/2024	
bit-lynk.com	3		15/04/2024	03/09/2024
bitaufy.com	1	1-22	23/03/2024	
bitcbel.com	4	4-1	02/12/2023	
bitcef.com	1	1-23	22/11/2023	
bitcefix.com	1	1-24	03/09/2023	03/09/2024
bitcfav.com	1		02/09/2023	03/09/2024
bitckeb.com	1	1-25	25/11/2023	
bitcker.com	1	1-26	25/11/2023	
bitckew.com	1	1-27	25/11/2023	
bitcland.com	1		02/09/2023	03/09/2024
bitcortexs.com	1	1-28	12/09/2023	13/09/2024
bitcotrend.com	1	1-29	12/12/2023	
bitcpop.com	1	1-30	25/11/2023	
bitcret.com	1	1-31	25/11/2023	
bitcusd.com	1	1-32	16/01/2024	
bitfires.org	3		25/10/2023	03/09/2024

bitflexup.com	3		01/12/2023	
bitfoxie.com	3		17/05/2024	03/09/2024
bitgalas.com	2	2-7	13/11/2023	
bitgetgift.com	1	1-33	03/04/2024	
bithorizen.com	1	1-34	07/09/2023	08/09/2024
bitmantos.com	2	2-8	21/11/2023	
bitmegahub.com	1	1-35	11/04/2024	
bitmohar.com	3		06/09/2023	03/09/2024
bitnase.com	1	1-36	17/01/2024	
bitnotex.com	1	1-37	27/10/2023	
bitoneu.com	1	1-38	26/02/2024	
bitrapay.com	3	3-5	22/11/2023	12/09/2024
bitrogex.com	1	1-39	08/10/2023	09/10/2024
bitruecoins.com	2	2-9	26/09/2023	27/09/2024
bits-wave.com	3	3-6	16/02/2024	
bitsplace.org	3	3-7	02/08/2024	
bitstader.com	1	1-40	16/02/2024	
bitstam.com	2	2-10	18/12/2023	
bitvevo.com	4	4-2	23/08/2023	
bitxbloom.com	1	1-41	06/02/2024	
bitxep.com	1	1-42	11/12/2023	
bitxnix.com	1	1-43	12/02/2024	
bixwen.com	4	4-3	13/12/2023	
blasbix.com	3	3-8	06/06/2024	
blocksjet.com	1	1-44	23/05/2024	
blocxkex.com	1	1-45	27/03/2024	
blogpx.com	1	1-46	15/12/2022	
btconmy.com	1	1-47	06/03/2024	
bufbit.com	2	2-11	21/11/2023	
bumtex.com	1	1-48	12/10/2023	
buytonbit.com	3		11/05/2024	03/09/2024
byfonex.com	1	1-49	20/09/2023	21/09/2024
byltex.com	2	2-12	03/09/2023	03/09/2024
bynxi.com	2	2-13	17/09/2023	18/09/2024
byte-flowx.com	3		09/11/2023	03/09/2024
bytly.site	3	3-9	15/07/2024	
bytwax.com	1	1-50	12/02/2024	
capcoinx.com	3	3-10	03/06/2024	
cashbitx.com	4	4-4	24/01/2024	
cashprimecoin.com	3		05/06/2024	03/09/2024
catxdex.com	1	1-51	23/07/2024	
caweu.com	2	2-14	28/06/2024	
chaingat.com	4	4-5	20/01/2024	
chaingex.com	2		02/09/2023	03/09/2024
chaingup.com	2	2-15	12/02/2024	
chainscade.com	1	1-52	08/12/2023	
chaintof.com	1	1-53	29/12/2023	
chainxlot.com	1	1-54	21/04/2024	
clezix.com	1	1-55	14/10/2023	
coehex.com	1	1-56	28/11/2023	
coezex.com	1	1-57	09/03/2024	
coin-oraclex.com	3		09/11/2023	03/09/2024
coin2bit.site	4	4-6	22/05/2024	
coinbitox.com	3	3-11	19/07/2024	
coincatalysthub.com	1	1-58	30/11/2023	
coinlex.us	1	1-59	05/10/2023	05/10/2024
coinport.live	1	1-60	02/07/2024	
coinsblocks.com	1	1-61	21/04/2024	
coinsbybase.com	3	3-12	06/08/2024	
coinshedge.com	1	1-62	28/11/2023	
coinstak.com	1	1-63	19/12/2023	
cointrustsbase.com	3		16/10/2023	03/09/2024
coinupfix.com	3	3-13	19/02/2024	
coinventus.com	2	2-16	26/09/2023	27/09/2024
coinwalletbase.com	3	3-14	09/12/2023	12/09/2024
coinxluxe.com	1	1-64	13/01/2024	
coinzentro.com	2	2-17	06/09/2023	07/09/2024
coinzotek.com	1	1-65	18/02/2024	
conbitex.com	2	2-18	23/11/2023	
coolchange.pics	1	1-66	28/02/2024	
copraex.com	2		02/09/2023	03/09/2024
cortexcoins.com	2		02/09/2023	03/09/2024
criptix.space	1	1-67	29/01/2024	
cromulentbit.com	3	3-15	27/06/2024	
cronoswap.com	1	1-68	04/01/2024	
cryp-sale.com	3		06/10/2023	03/09/2024
crypdax.com	3	3-16	02/07/2024	
crypsam.com	3	3-17	28/03/2024	23/09/2024
cryptalchange.org	2	2-19	08/09/2023	08/09/2024
crypterians.com	1	1-69	03/09/2023	03/09/2024
cryptiones.com	1	1-70	11/12/2023	
crypto-a-i.com	3	3-18	08/02/2024	
cryptofamas.com	1	1-71	16/12/2023	
cryptogetx.com	3		17/09/2023	03/09/2024

cryptohubcoin.com	4	4-7	20/11/2023	
cryptomeyou.com	1	1-72	04/09/2023	05/09/2024
cryptomuss.com	2	2-20	26/11/2023	
cryptoprimecash.com	3	3-19	03/08/2024	
cryptowos.com	2	2-21	09/09/2023	10/09/2024
cryptowyx.com	4	4-8	20/01/2024	
cryptox-hub.com	1	1-73	06/02/2024	
cryptoxbeast.com	1	1-74	05/12/2023	
crypzar.com	3	3-20	13/06/2024	
crypze.com	1	1-75	07/05/2024	
cyberbitex.com	1	1-76	05/12/2023	
dayxtrade.com	1	1-77	05/09/2023	06/09/2024
dayxtrades.com	1	1-78	13/09/2023	14/09/2024
definton.com	4	4-9	29/12/2023	
devourex.com	1	1-79	18/12/2023	
deweu.com	1	1-80	18/07/2024	
dex-mall.com	3		15/11/2023	03/09/2024
dexenos.com	1	1-81	09/09/2023	10/09/2024
dexhuobi.io	3		15/06/2024	
dexinfinity.com	3		20/06/2024	
dextier.com	2	2-22	28/06/2024	
dexupspace.com	2	2-23	06/09/2023	07/09/2024
dexzond.com	1	1-82	27/12/2023	
digitalrpx.com	1	1-83	05/12/2023	
dogespase.com	1	1-84	23/03/2024	
dolbon.com	1	1-85	24/10/2023	
dolebix.com	1	1-86	12/12/2023	
doletex.com	1	1-87	26/10/2023	
doxespace.com	1	1-88	12/04/2024	
dreamzex.com	1	1-89	21/10/2023	
dropdogex.com	1	1-90	23/03/2024	
dytebix.com	1	1-91	25/11/2023	
eburedex.com	1	1-92	12/09/2023	13/09/2024
elbitx.com	1	1-93	01/11/2023	
elonuts.com	1	1-94	22/12/2023	
emunahvault.com	1	1-95	08/04/2024	
emweu.com	1	1-96	17/08/2024	
etherxspark.com	1	1-97	03/01/2024	
ethxfy.com	1	1-98	12/02/2024	
etodrops.com	1	1-99	04/03/2024	
excrynex.com	1	1-100	16/09/2023	17/09/2024
exfitex.com	1	1-101	02/09/2023	03/09/2024
exohoods.com	1	1-102	30/11/2023	
exounions.com	1	1-103	05/12/2023	
explosetrade.com	4	4-10	13/12/2023	
exrnchains.com	1	1-104	05/12/2023	
extezones.com	1	1-105	18/11/2023	
exybase.com	1	1-106	13/10/2023	
fasterdex.com	2	2-24	01/06/2024	
fasxdex.com	1	1-107	16/07/2024	
fewdaotrade.com	1	1-108	16/12/2023	
fezau.com	1	1-109	15/01/2022	
finanecoin.com	1	1-110	11/12/2023	
finsconvex.com	1	1-111	05/09/2023	06/09/2024
flavtex.com	1	1-112	11/10/2023	
fleqcoin.com	1	1-113	15/09/2023	16/09/2024
float-token.com	2	2-25	23/07/2024	
foldyx.com	2	2-26	05/09/2023	06/09/2024
forbes-coins.com	2	2-27	28/10/2023	
forgebitx.com	1	1-114	24/01/2024	
forusax.com	1	1-115	11/09/2023	12/09/2024
frebtc.com	2	2-28	06/11/2023	
futureoxup.com	1	1-116	11/09/2023	12/09/2024
gadadex.com	1	1-117	12/04/2024	
ganbitc.com	4	4-11	05/10/2023	05/10/2024
gashexer.com	1		02/09/2023	03/09/2024
gatesword.com	1	1-118	16/12/2023	
gatexnet.com	1	1-119	30/05/2024	
geetwix.com	1	1-120	15/09/2023	16/09/2024
gerinbit.com	1	1-121	07/11/2023	
gevetex.com	1	1-122	15/09/2023	16/09/2024
giftdefi.com	2	2-29	30/10/2023	
glidafin.com	1	1-123	13/12/2023	
gludaps.com	2	2-30	11/08/2024	
gradebitex.org	3		05/09/2023	03/09/2024
grespace.com	1	1-124	28/11/2023	
grpcomp.com	1	1-125	11/11/2023	
gytesed.com	1	1-126	26/12/2023	
hadderbit.com	1	1-127	29/10/2023	
haederbit.com	1	1-128	16/09/2023	16/09/2024
haeterbit.com	1	1-129	30/09/2023	01/10/2024
haiderbit.com	1	1-130	27/10/2023	
hasterbit.com	1	1-131	11/12/2023	
hatterbit.com	1	1-132	11/10/2023	

haydderbit.com	1	1-133	20/10/2023	
hayderbid.com	1	1-134	18/11/2023	
heiterbit.com	1	1-135	31/10/2023	
heydderbit.com	1	1-136	24/10/2023	
heyderbid.com	1	1-137	18/11/2023	
hobetex.com	1	1-138	24/11/2023	
hollbit.com	2	2-31	25/09/2023	26/09/2024
holpeks.com	1	1-139	06/11/2023	
hordcoin.com	1	1-140	04/10/2023	05/10/2024
hotxdex.com	1	1-141	16/07/2024	
howzex.com	1	1-142	15/02/2024	
hydtex.com	1	1-143	03/11/2023	
ibitsi.com	2	2-32	13/12/2023	
iriscoin-ex.com	3	3-21	22/02/2024	
kanotex.com	1	1-144	13/10/2023	
khaderbit.com	1	1-145	04/09/2023	05/09/2024
khaiterbit.com	1	1-146	08/11/2023	
kopraex.com	2	2-33	05/10/2023	06/10/2024
krakendefi.com	2	2-34	30/10/2023	
kunydex.com	2	2-35	25/07/2024	
lanterex.com	1	1-147	14/09/2023	15/09/2024
laodaobit.com	3		29/04/2024	
lightexpro.com	1	1-148	08/05/2024	
limbobit.net	3		16/04/2024	03/09/2024
lite-dex.com	3		07/04/2024	03/09/2024
lixbycoin.com	1	1-149	01/12/2023	
lydianex.com	1	1-150	28/06/2024	
mababit.com	3		13/04/2024	
macxgates.com	2	2-36	01/12/2023	
mercbit.com	1	1-151	23/02/2024	
mestanex.com	1	1-152	24/10/2023	
metacoinswallet.com	3		06/09/2023	03/09/2024
metaprojectx.com	1	1-153	19/09/2023	20/09/2024
mexyce.com	2	2-37	04/01/2024	
modalitycoin.com	1	1-154	01/02/2024	
mooprex.com	2	2-38	10/09/2023	11/09/2024
mrbeastgive.com	1	1-155	13/03/2024	
mrbeastox.com	1	1-156	24/11/2023	
musthex.com	1	1-157	05/12/2023	
myaltcoins.org	3		21/08/2023	
mynexcoin.com	1	1-158	23/11/2023	
nasvayx.com	1	1-159	18/09/2023	19/09/2024
neardexer.com	1	1-160	11/06/2024	
nelenex.com	4	4-12	15/05/2024	
nelotex.com	1	1-161	15/10/2023	
nemodrop.com	1	1-162	12/04/2024	
nenarex.com	2	2-39	30/04/2024	
netlobit.com	3	3-22	24/06/2024	
nexodiums.com	1	1-163	26/09/2023	27/09/2024
nextcoinbank.com	3		08/11/2023	03/09/2024
noimex.com	1	1-164	30/12/2023	
nowesp.com	1	1-165	27/03/2024	
nowpacex.com	4	4-13	08/01/2024	
nowzex.com	1	1-166	04/02/2024	
nuschain.com	2	2-40	19/12/2023	
obabit.com	1	1-167	16/10/2023	
obeebit.com	1	1-168	16/10/2023	
obmanex.com	1	1-169	06/12/2023	
oktibit.com	1	1-170	05/09/2023	06/09/2024
olympcoin.io	2	2-41	31/03/2023	
onestarbtc.com	2	2-42	10/11/2023	
opemdix.com	1	1-171	18/11/2023	
opozon.com	1	1-172	02/06/2024	
orioncryptoexchange.com	1	1-173	18/06/2024	
osweu.com	1	1-174	21/07/2024	
ozauk.com	1	1-175	14/11/2022	
panosix.com	1	1-176	12/01/2024	
payspex.com	1	1-177	05/12/2023	
pepevivex.com	1	1-178	20/09/2023	21/09/2024
pepnex.com	1	1-179	24/10/2023	
pibitex.com	1	1-180	25/01/2024	
pitbitc.com	4	4-14	14/10/2023	
pixcryptos.com	1	1-181	12/12/2023	
polymatex.com	2	2-43	08/09/2023	09/09/2024
porezex.com	1	1-182	07/10/2023	08/10/2024
poxdrop.com	1	1-183	02/05/2024	
projectxpro.space	2	2-44	19/09/2023	20/09/2024
pyrexcoins.com	1	1-184	12/09/2023	13/09/2024
qexchain.com	1	1-185	06/09/2023	07/09/2024
qwitex.com	2	2-45	11/10/2023	
raderex.com	2	2-46	27/11/2023	
raisenexus.com	2		02/09/2023	03/09/2024
regbitex.com	2	2-47	25/04/2024	
remrex.com	1	1-186	17/10/2023	

remxbit.com	4	4-15	19/09/2023	19/09/2024
rexybit.com	3	3-23	03/07/2024	
richardbit.com	3	3-24	09/11/2023	
risewex.com	4	4-16	05/05/2024	
robycoin.com	1	1-187	21/12/2023	
roxycoin.com	2	2-48	06/09/2023	07/09/2024
safefex.com	1	1-188	04/12/2023	
safekes.com	1	1-189	24/01/2024	
saitabits.com	1	1-190	18/12/2023	
sallbit.com	4	4-17	08/05/2024	
salorex.com	2	2-49	26/10/2023	
savorbit.com	3	3-25	08/07/2024	
secoinbix.com	1	1-191	05/12/2023	
sedrops.com	1	1-192	27/01/2024	
sefava.com	1	1-193	24/01/2024	
siacoinex.com	2	2-50	17/10/2023	
skiletex.com	1	1-194	04/09/2023	05/09/2024
skobrex.com	1	1-195	12/09/2023	13/09/2024
skoxdex.com	1	1-196	06/07/2024	
skytore.com	1	1-197	26/06/2024	
sobrecoin.com	3		31/08/2023	
socbitex.com	1	1-198	05/12/2023	
solenexs.com	2		02/09/2023	03/09/2024
soltkey.com	2	2-51	16/01/2024	
sonarexs.com	2	2-52	28/09/2023	29/09/2024
sopecex.com	1	1-199	01/01/2024	
sosbex.com	1	1-200	04/09/2023	05/09/2024
soweu.com	2	2-53	06/04/2024	
spacemuske.com	1	1-201	05/12/2023	
spacepex.com	1	1-202	31/10/2023	
spacetux.com	1	1-203	03/09/2023	04/09/2024
spacevex.com	1	1-204	25/11/2023	
spacexplace.com	1	1-205	03/10/2023	04/10/2024
spaceyam.com	1	1-206	30/11/2023	
spasetex.com	1	1-207	05/01/2024	
spatedex.com	1	1-208	21/07/2024	
spectralcoin.com	1	1-209	25/09/2023	26/09/2024
sponbix.com	1	1-210	05/12/2023	
spoxdex.com	1	1-211	16/07/2024	
srenex.com	1	1-212	10/10/2023	10/10/2024
stakecoinup.com	3	3-26	23/08/2024	
stakexzero.com	1	1-213	11/03/2024	
starkalt.com	1	1-214	30/04/2024	
starkxdrop.com	1	1-215	11/03/2024	
stellar-x-change.com	3		04/10/2023	03/09/2024
stellarcoin.pro	1		03/07/2024	
stellarxo.com	1	1-216	11/11/2023	
stocks-dex.com	3	3-27	08/02/2024	
stronbex.com	2	2-54	07/11/2023	
sun-dex.com	3	3-28	28/05/2024	
supexer.com	1	1-217	28/03/2024	
supxdex.com	1	1-218	16/07/2024	
svenex.com	1	1-219	30/09/2023	01/10/2024
swapxtrade.com	1	1-220	21/02/2024	
tardexs.com	1	1-221	28/11/2023	
taybitc.com	1	1-222	27/10/2023	
tedrops.com	1	1-223	06/03/2024	
tedwex.com	1	1-224	07/06/2024	
temebax.com	1	1-225	12/03/2024	
tensedex.com	2	2-55	04/02/2024	
teondex.com	2	2-56	21/07/2024	
terdebix.com	1	1-226	12/12/2023	
teronex.com	1	1-227	28/12/2023	
tesdrops.com	1	1-228	04/03/2024	
thronexs.com	1	1-229	16/09/2023	17/09/2024
to-exchange.org	3		08/02/2023	03/09/2024
togedex.com	1	1-230	07/06/2024	
tokenmak.com	1	1-231	02/09/2023	03/09/2024
tokenomys.com	1	1-232	05/12/2023	
topwebtc.com	2	2-57	06/11/2023	
toranex.com	4	4-18	25/04/2024	
tradeplex.io	1	1-233	02/04/2024	
trader-global.com	2	2-58	28/06/2024	
tradexturbo.com	3	3-29	30/10/2023	03/10/2024
trongex.com	1	1-234	30/12/2023	
truebit.biz	2	2-59	04/12/2023	
trustcoinswallet.com	3		16/08/2023	
trustexer.com	1	1-235	11/09/2023	12/09/2024
trusts-coin.com	2	2-60	17/10/2023	
twitbex.com	2	2-61	25/09/2023	25/09/2024
typesuncoin.com	1	1-236	06/03/2024	
unezex.com	1	1-237	09/02/2024	
unixloop.com	1	1-238	17/10/2023	
upbitx.net	2	2-62	17/07/2024	

upecex.com	1	1-239	03/01/2024	
usaxmoon.com	1	1-240	12/12/2023	
usetradex.com	3	3-30	04/07/2024	
usweu.com	1	1-241	19/05/2024	
uvines.com	1	1-242	23/03/2019	
valterex.com	1	1-243	29/06/2024	
vapyxer.com	4	4-19	19/11/2023	
vaultexa.com	1	1-244	27/05/2024	
vazedex.com	1	1-245	13/03/2024	
veko24.com	1	1-246	30/10/2023	
vergexs.com	1	1-247	26/10/2023	
vexpump.com	1	1-248	12/04/2024	
vortexswift.com	1	1-249	26/06/2024	
voxsu.com	1	1-250	14/11/2022	
vozdex.com	1	1-251	06/07/2024	
walkenbit.com	1	1-252	29/09/2023	30/09/2024
waltochain.com	2	2-63	14/09/2023	15/09/2024
wangbitc.com	4	4-20	05/10/2023	06/10/2024
wenanzo.com	1	1-253	05/11/2023	
wespacex.com	1	1-254	23/11/2023	
wexne.com	1	1-255	13/11/2022	
wezudex.com	1	1-256	13/03/2024	
whitecoinex.com	2	2-64	26/10/2023	
wifbit.com	2	2-65	22/04/2024	
wirextokens.com	1		02/09/2023	03/09/2024
wongbit.com	4	4-21	14/10/2023	
worktravelrecruting.de	1	1-257	26/05/2024	
wryzitex.com	1	1-258	28/11/2023	
xbitvault.com	4	4-22	25/04/2024	
xdropex.com	1	1-259	02/07/2024	
xelonmex.com	1	1-260	16/09/2023	17/09/2024
xmocrypto.com	1	1-261	17/12/2023	
xnovadex.com	1	1-262	10/08/2024	
xodropex.com	1	1-263	30/05/2024	
yaseibit.com	1	1-264	06/09/2023	07/09/2024
yazeibit.com	1	1-265	03/09/2023	03/09/2024
yeblance.com	1	1-266	02/09/2023	03/09/2024
yeplance.com	1	1-267	11/10/2023	
zaglex.com	1	1-268	20/11/2023	
zandonex.com	1	1-269	09/03/2024	
zelotex.com	1	1-270	23/10/2023	
zelotix.com	1	1-271	27/10/2023	
zespacex.com	1	1-272	27/11/2023	
zexdropex.com	1	1-273	02/05/2024	
zwapexe.com	1	1-274	01/12/2023	
zylobit.com	1	1-275	02/07/2024	